Structure and Representation of Ecological Data to Support Knowledge Discovery: a Case Study with Bioacoustic Data

by

Frank Pouw

B. Sc., Simon Fraser University, 1993

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN ENVIRONMENTAL SCIENCE

Thesis Examining Committee:

Mila Kwiatkowska (PhD) Associate Professor in the Department of Computing Science and Thesis Supervisor, Thompson Rivers University

Tom Dickinson (PhD) Professor and Dean of Science and Committee member, Thompson Rivers University

Matt Reudink (PhD) Associate Professor in the Department of Biological Sciences and Committee member, Thompson Rivers University

Mario Nascimento (PhD) Professor in the Department of Computing Science, University of Alberta

May, 2019

Thompson Rivers University

Thesis Supervisor: Associate Professor Mila Kwiatkowska

Abstract

Bird communities have long been surveyed as key indicators of ecosystem health and biodiversity. Adoption of Autonomous Recording Units (ARUs) to perform avian surveys has shifted the burden of species recognition from “birders” in the field, to “listeners” who review the ARU recordings at a later time. The number of recordings ARUs can produce has created a need to process large amounts of data. Although much research is devoted to fully automating the recognition process, expert humans are still required when entire bird communities must be identified. A framework for a Decision Support System (DSS) is presented which would assist listeners by suggesting likely species. A unique feature of the DSS is the consideration of the recording “context” of time, location and habitat as well as the bioacoustic features to match unknown vocalizations with reference species.

In this thesis a data warehouse was built for an existing set of bioacoustic research data as a first–step to creating the DSS. The data set was from ARU deployments in the Lower Athabasca Region of Alberta, Canada. The Knowledge Discovery in Databases (KDD) and Dimensional Design Process protocols were used as guides to build a Kimball–style data warehouse. Data housed in the data warehouse included field data, data derived from GIS analysis, fuzzy logic memberships and symbolic representation of bioacoustic recording using the Piecewise Aggregate Approximation and Symbolic Aggregate approXimation (PAA/SAX). Examples of how missing and erroneous data were detected and processed are given. The sources of uncertainty inherent in ecological data are discussed and fuzzy logic is demonstrated as a soft–computing technique to accommodate this data.

Data warehouses are commonly used for business applications but are very applicable for ecological data. As most instructions on building data warehouse are for business data, this thesis is offered as an example for ecologists interested in moving their data to a data warehouse. This thesis presents a case–study of how a data warehouse can be constructed for existing ecological data, whether as part of a DSS or a tool for viewing research data.

Keywords: bioacoustics, decision support system, data warehouse, fuzzy logic, birds, autonomous recording units, Piecewise Aggregate Approximation, Symbolic Aggregate Approximation

Contents

Acknowledgements

I would like to express my sincere thanks to my thesis advisor Dr. Mila Kwiatkowska and my committee members, Dr. Tom Dickinson and Dr. Matt Reudink. Their guidance, encouragement and patience greatly assisted me through this rewarding process.

I must also give my warmest thanks to Dr. Erin Bayne who generously allowed to to use his data, without which this thesis would not have been possible.

Thanks as well go to my wife who's constant support and proof–reading skills are greatly appreciated.

Lastly, I must acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada (NSERC).

List of Figures

List of Tables

Nomenclature

biophonic sound Sounds made by organisms

birder an expert humans trained to identify birds by sound and sight and employed to conduct avian field surveys

context The conditions under which an acoustic event was observed or recorded. Includes the time of day, time of season, geographic location and habitat.

crisp numeric values which have no uncertainty

data mart A small dimensional data base modeled to a specific business process. A component of the data warehouse design introduced by Ralph Kimball

data warehouse A non-normalized database architecture optimized to allow data analysis of unchangine data.

dimension table A table of a Dimensional Database which is used to store descriptive attribues which can be used as a grouping parameter

dimensional database model A database architecture comprised of a central fact table to stores numeric attributes and which is joined to dimension tables that store descriptive attributes

fact table A table of a dimensional database which is used to store numerical attribues

FIS see *fuzzy inference system*

fuzzy inference system A model based on fuzzy logic and composed of fuzzy membership functions and fuzzy rules.

geophonic sound Sounds made by non-organic natural processes

KDD See *Knowledge Discovery in Databases*

Knowledge Discovery in Databases The identification of non-trivial relationships within a large set of data

listener expert human capable of identify birds by sound and employed to make species determinations from bioacoustic recordings

normalization rules A sequence of conditions implemented to reduce redundancy in a transactional database

operational database See *transactional database*.

recording profile The acoustic, temporal and spatial characteristics associated with a recorded acoustic event.

shapelet a short patterns that define a class within a time series

species profile The acoustic, temporal and spatial characteristics associated with a particular species.

transactional database A normalized database architecture optimized to maintain the accurace storage of dynamic data.
Also called a operational database.

decision support system A computer based tool which analyzes data for the purpose of assisting a person to make an informed choice

Chapter 1

Introduction

Autonomous Recording Units in Avian Surveys

Avian surveys have long been used to study individual bird species, to estimate ecosystem biodiversity and as an indicator of environmental health because birds are widespread across many habitats, are sensitive to environmental change and are relatively easy to survey (Brandes, 2008; Gregory and Strien, 2010; Chambert et al., 2018). The traditional method of conducting avian surveys relies on expert human "birders" who locate themselves at predetermined locations and record all the birds they can identify by sound and sight within the time and distance stipulated by a standardized protocol such as the North American Breeding Bird Survey (BBS) (Sauer et al., 2013). These surveys are limited by the availability of trained birders and the logistics of moving birders to survey sites during the time birds are most active. These restrictions force researchers to choose between the number of sites which can be surveyed and the intensity to which each site is sampled. This has lead to a need to develop techniques which improve avian surveys in order to improve the information they provide to science and conservation efforts (Brandes, 2008). One such technique is the adoption of Autonomous Recording Units (ARUs).

ARUs are robust, computer–controlled acoustic recorders which can be left in the field to record at a preset intervals. They have been shown to have the potential to expand the capacity of avian monitoring (Hutto and J. Stutzman, 2009; Rempel et al., 2005; Brandes, 2008; La and Nudds, 2016) chiefly because they decouple the occurrence of an

acoustic event (e.g.: a bird song) from the analysis of that event (e.g.: the species determination by an expert birder). In contrast to the traditional bird survey technique, ARUs can be deployed and retrieved by non-specialized field technicians, after which the recordings are analyzed by experts human "listeners." While this has greatly increased the acquisition of the avian survey recordings, the method has introduced two problems. The analysis of large amounts of bioacoustic recordings (Zhang et al., 2016) and the dissociation of an acoustic event from the conditions under which it occurred.

Ideally, identifying bird species from ARU recordings could be completely automated, and a great deal of work is being done in this field. However, omnidirectional field recordings present many challenges to computer processing which are trivial to human observers including duets, choruses of overlapping songs and species with regional dialects and species with very large repertoires and improvisational songs (Brandes, 2008) and are prone to errors (Chambert et al., 2018). Consequently most projects concentrate on the identification of one or a few species (Fagerlund, 2004) and most still rely on expert–human confirmation (Shonfield and Bayne, 2017b; Darras et al., 2017). Therefore, procedures to identify all species on a recording still requires the judgment of expert human listeners (Wimmer et al., 2013; Keen et al., 2014).

The dissociation of acoustic events from their environment is a less obvious problem for ARU processing. While listeners have the advantage of repeatedly listening to difficult recordings, visualizing sonograms, accessing reference recordings and the advice of other listeners, they must make their species determination based solely on the bioacoustic signal. In contrast, birders conducting traditional field surveys benefit from experiencing the temporal and spatial conditions under which a vocalization occurred. This knowledge of the environment, hereafter called "context", includes time–of–day, time–of–season, geographic location and habitat. With an awareness of context and knowledge of behaviour, migration patterns, distribution and habitat preference of candidate species, an experienced birder can compose a list (perhaps subconsciously) of birds likely to be encountered. For example, Dusky Flycatcher (*Empidonax oberholseri Phillips, AR, 1939*) and Hammond's Flycatcher (*Empidonax hammondii* (Xántus de Vesey, 1858)) have very similar songs but *E. oberholseri* prefers shrubby open sites while *E. hammondii* prefers sites with closed canopies (Mannan, 1984; Sedgwick, 1975). A birder may instinctively feel that, "this seems like Dusky habitat."

The effort to keep pace with the number of recordings which ARUs can create is significant (Agranat, 2009; Shonfield and Bayne, 2017b; Zhang et al., 2016) . Since fully

automated recognition of all species in an avian community is still beyond the scope of current computer systems (Truskinger et al., 2015; Chambert et al., 2018; Darras et al., 2017); methods of assisting human–based processing are necessary to manage the increasing number of ARU recordings being acquired (Charif and Pitzrick, 2008; Shonfield and Bayne, 2017b). To meet this need, Decision Support Systems (DSS) for bioacoustic processing are being developed (Truskinger et al., 2015, 2011) but the systems so far proposed rely only on identifying similar patterns within the acoustic signal. This thesis presents the framework for a software–based DSS which would use both bioacoustic and context data to provide listeners with a list of suggestions for species likely to have made a vocalization. To the author's knowledge, including context in a DSS for bioacoustic recognition is a novel approach. The complete development of a DSS is beyond the scope of this thesis, what is presented here is the first–step towards its development: a data warehouse to manage the existing bioacoustic data.

Data Warehouse for the Purposes of Knowledge Discovery

The techniques to match acoustic and context data with reference recordings and life history knowledge fall within the discipline of data mining, also called Knowledge Discovery in Databases (KDD). Data mining and data warehouses have been called the "architectural foundation of a decision support system" (Inmon, 1996) because a properly developed data warehouse promotes efficient access to reliable data by collecting, cleaning and grouping data from a variety of sources in a standardized format (Hinton, 2006). The development of a data warehouse extends beyond its utility in the proposed DSS. Loading data into a data warehouse is a necessary step in the adopting of a "Big–data culture" where diverse data is combined, re–purposed and from which patterns can be discovered which in turn can motivate new lines of research (Hampton et al., 2013; Palmer et al., 2005).

Research Objective

The focus of this research was to construct a data warehouse for existing bioacoustic data and context data in order to facilitate knowledge discovery. The data warehouse was designed as a component of a Decision Support System to assist human listeners processing bioacoustic recordings. As data warehouses were originally developed and

have been used primarily for business applications, particular attention was given to the unique challenges presented by ecological data. It is hoped that the findings of this thesis will provide a useful example for other ecologists wishing to move their data into a data warehouse.

The remainder of this thesis is structured as follows: Chapter 2 discusses data management in ecology and describes the structure of a data warehouse. Additionally, several techniques not widely used in ecology are introduced. Chapter 3 presents the steps taken to process the data and the creation of the data warehouses. Conclusions are presented in Chapter 4.

Chapter 2

Discovering Knowledge in Ecological Data

Introduction

In this chapter an outline of the Decision Support System (DSS) framework, proposed in Chapter 1, is presented so the reader can understand the function the data warehouse was designed to serve. This is followed by a description of the data warehouse structure, as developed by Ralph Kimball (1996), with comparisons to the two most common ways ecological data is stored: spreadsheets and transactional databases. Two unique challenges with bioacoustic data are then discussed. Firstly, the representation of acoustic files, where the techniques of Piecewise Aggregate Approximation (PAA) (Yi and Faloutsos, 2000; Keogh et al., 2000) and Symbolic Aggregate Approximation (PAA/SAX) (Lin et al., 2003) are illustrated with an example. Secondly, the sources of uncertainty in ecological data are discussed and Fuzzy Logic (Zadeh, 1965) is offered as one technique suitable to represent imprecise ecological attributes. Examples are given to illustrate the use and derivation of fuzzy sets and how they can be employed in a fuzzy inference system.

A Decision Support System for Bioacoustic Processing

The proposed DSS would contribute to a listener–based processing system by suggesting a list of likely species for each acoustic event, based on characteristics of the acoustic signal and the recording context. Figure 2.1 presents the framework for the DSS. Acknowledging that there are many challenges, such as isolating individual

vocalizations from a bioacoustic recording, conceptually the steps of the DSS are as follows:

First, data are collected as inputs. Two types of data are used by the DSS, field data and reference data. Field data includes bioacoustic field recordings, associated metadata and the recording context. Reference data includes acoustic reference recordings and life history knowledge for each species in the study area.

Second, input data are codified and organized into species profiles and a recording profiles. A species profile, comprised of acoustic features extracted from reference recordings and habitat, range and behaviour characteristics, is created for each species in the study area. A recording profile comprised of extracted acoustic features, spatial and temporal recording context is created for each acoustic event on the bioacoustic recordings.

Third, a comparison is made between the recording profile of each unknown acoustic event and each species profile. The strength of a match is based on both the similarity of acoustic features and the equivalence of the recording context with the habitat, range and behaviour.

Fourth, for each acoustic event, a ranked list of potential species is presented to listeners based on the pattern matching.

Fifth, for each acoustic event a species determination is made by the listener based on their appraisal of the acoustic event and the suggestions made by the DSS. This assessment is recorded in the data repository and could be used to refine the species profile made in step two.

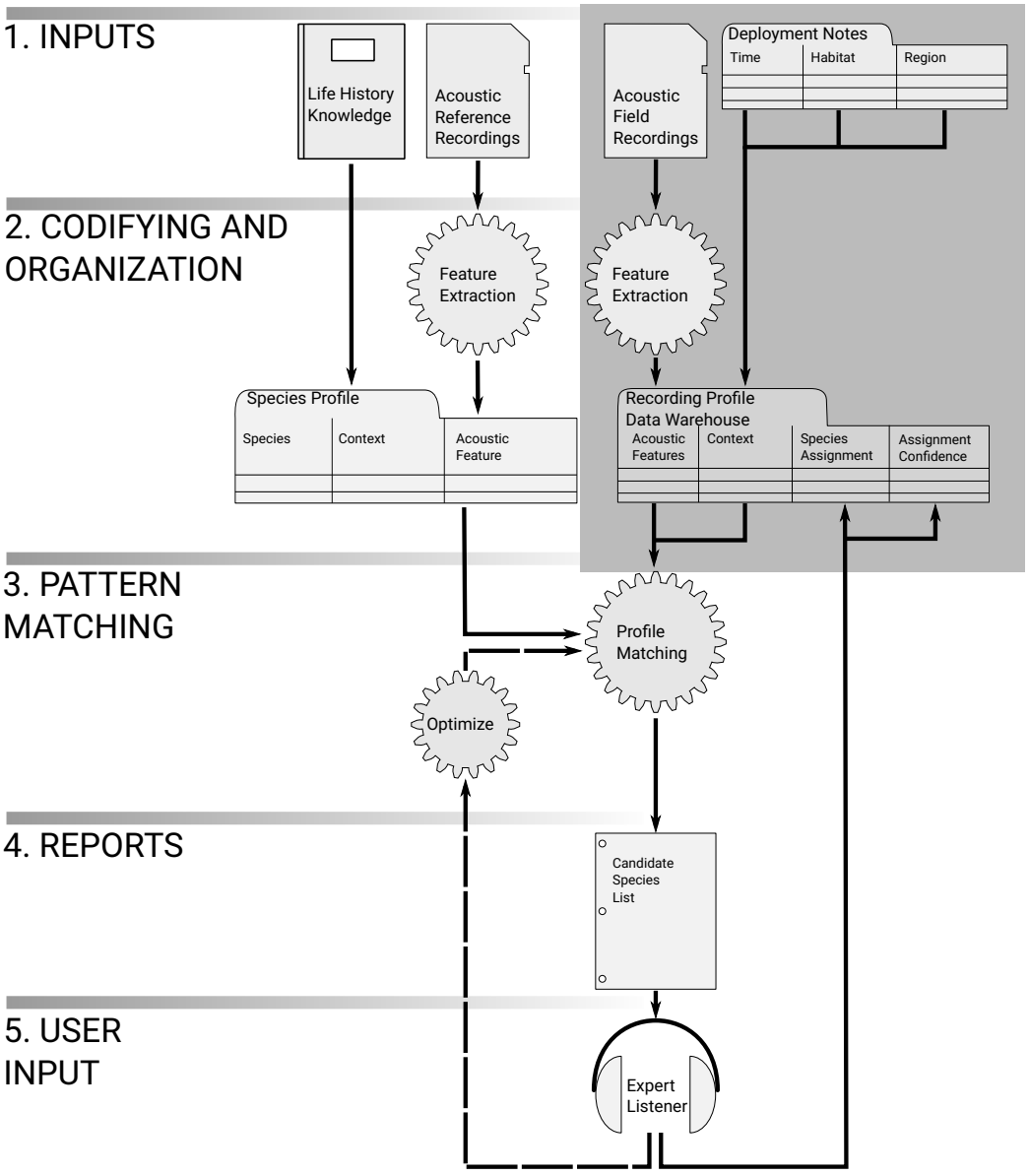


Figure 2.1: Components of a Decision Support System proposed to facilitate species identification by humans from bioacoustic recordings. The grey box indicates the focus of this research which was the creation of a data warehouse for bioacoustic data.

The reference patterns for each candidate species would be derived from reference recordings combined with knowledge of distribution and life–history information.

While some reference knowledge can be gathered from sources such as the Macaulay Library at the Cornell Lab of Ornithology (www.birds.cornell.edu) and the Bird Studies Canada Breeding Bird Atlas (www.birdscanada.org), the DSS can also learn from identified acoustic events on ARU recordings.

Storage Structures for Ecological Data

Knowledge in data takes the form of patterns which indicate relationships between system components. The process by which knowledge is found in databases is called Knowledge Discovery in Database (KDD), also called data mining.

Although it is possible to mine data from many data structures, the right data structure will improve the chances of success (Inmon, 1996). Often ecological data is stored in minimally structured formats such as flat files (e.g.: spreadsheets) which are not conducive to knowledge discovery (Madin et al., 2007), do not adequately constrain data to enforce data integrity (Jones et al., 2006) and do not allow researchers to easily view data in novel ways.

A superior data management tool is the relational database (such as Oracle 12c, MySQL and MS Access) which stores data in linked tables (called relations) and can both constrain data input and allow data to be grouped and re–arranged. Two types of database architectures, transactional databases and data warehouses, are suitable for the management of ecological data.

The transactional database is designed to accurately manage changing data. To prohibit erroneous data from being entered, the database is built to comply with normalization rules which preclude redundancy within data rows (i.e., records) and between columns (i.e., attributes) (Hillyer, 2005). As a consequence, a normalized database will have many tables, each containing data related to one specific subject. Although this design will enforce data integrity, the large number of tables and relationships increases the complexity of extracting data.

For example, a transactional database designed to store data from vegetation surveys is shown in Figure 2.2 and is comprised of seven tables and seven relationships. With this structure, instead of recording the name of a particular plant in the UNDERSTORY table each time it is found, a link is made to that species record in the PLANT_SPECIES table.

Here the name is recorded only once, thus eliminating the possibility of misspelled entries or the use of outdated names which could occur if redundancy were allowed.

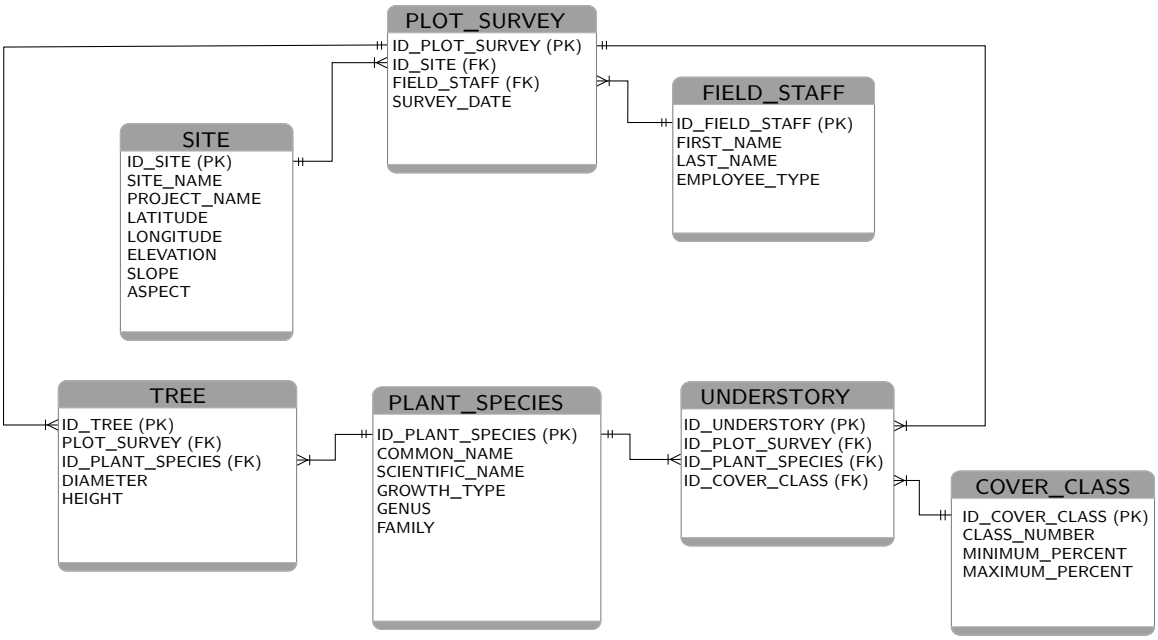


Figure 2.2: Example of a normalized transactional database structure designed to record vegetation survey data, showing how information for each subject is recorded only once in its own table.

While transactional databases are appropriate for the collection, validation and correction of research data, their complex structure makes data extraction difficult (Moody and Kortink, 2000). In the above example, to generate a report shown in Table 2.1, six of the tables must be queried, which illustrates the trade–off between enforcing data integrity and the ease and efficiency of querying data.

Table 2.1: A typical report generated from the example vegetation database.

Site	**Date**	**Species**	**Height**	**Diameter**	**Cover (Min)**	**Cover (Max)**
Rd_01	2017/06/24	*B. nana*	6	12	null	null
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Another database architecture, the data warehouse, was designed to facilitates data queries for online analytical processing, an approach to analyzing business data (Inmon, 1996; Thuraisingham, 1997). Although designed for business managers seeking to make

decisions based on trends within business data, data warehouses are equally suited for ecologists seeking to discover patterns within ecological data. Information queried from a data warehouse can be passed to statistical programs, Geographic Information Systems and data mining software for further analysis.

This ease of data extraction is achieved by allowing redundancy in the data (Sen and Sinha, 2005). In the aforementioned vegetation database, the MINIMUM_PERCENT and MAXIMUM_PERCENT of cover (the threshold of each cover class) could be included in each record for a species found in a vegetation plot, eliminating the need for the COVER_CLASS table and thus simplifying queries. But, unlike the redundancy occurring in a flat file (such as a spreadsheet), the possibility of errors is avoided by closely managing the sources of data imported into the data warehouse.

Data warehouses function as read–only sources (Moody and Kortink, 2000) of unchanging data (Gupta, 1997; El-Sappagh et al., 2011). In business applications this would be historic sales data, in ecological applications this would be data collected from field observations. The two most common data warehouse designs are the enterprise data warehouse proposed by Inmon and Kelley (1993) and the collection of smaller databases, called “data marts”, proposed by Kimball (1996) (Breslin, 2004; Sen and Sinha, 2005). Both were developed to assist business managers to make decisions based on trends in sales data, but differ in architecture and implementation. These are commonly called the Inmon model and the Kimball model.

The Inmon model employs a top–down approach where a single, monolithic data warehouse is created first, from which smaller databases are extracted for the needs of individual departments. This model requires an advanced level data modelling expertise and preexisting knowledge of analytical requirements (Breslin, 2004). Because ecologists often manage their own data and must accommodate unanticipated research directions, the Inmon model is not suitable and will not be discussed further.

The Kimball model uses an approach which is both bottom–up and top–down, by creating several data marts, each of which is modelled for a specific business process (or research activity). Conformity is maintained between data marts, allowing them to be used together. The simplicity of each data mart, and the ability to add new data marts as different types of data are collected, provides the simplicity and expandability required by ecological researchers. A description of the Kimball data warehouse model follows.

Data marts are built with the “Dimensional Database Model” architecture comprised of two types of tables: a Fact Table which stores all quantifiable metrics and Dimension

Tables which contain qualitative attributes (Kimball et al., 2002; Moody and Kortink, 2000). The granularity of the facts (i.e., the smallest unit in which a record can be divided) should be small in order to produce a data warehouse adaptable to *ad hoc* queries (Kimball et al., 2008). Metrics in the fact table are grouped based on attributes in the dimension tables.

A typical data mart contains a single fact table joined to several dimension tables. For example, Figure 2.3 illustrates how data from the transactional vegetation database (Figure 2.2) could be stored in a data warehouse. Numeric metrics like SLOPE and HEIGHT are stored in the Fact Table while descriptive categories like SITE_NAME and GENUS are stored in Dimension Tables. Also shown is the option to store commonly derived data, such as AVERAGE_PERCENT_COVER which is calculated from the stored MINIMUM_PERCENT and MAXIMUM_PERCENT values.

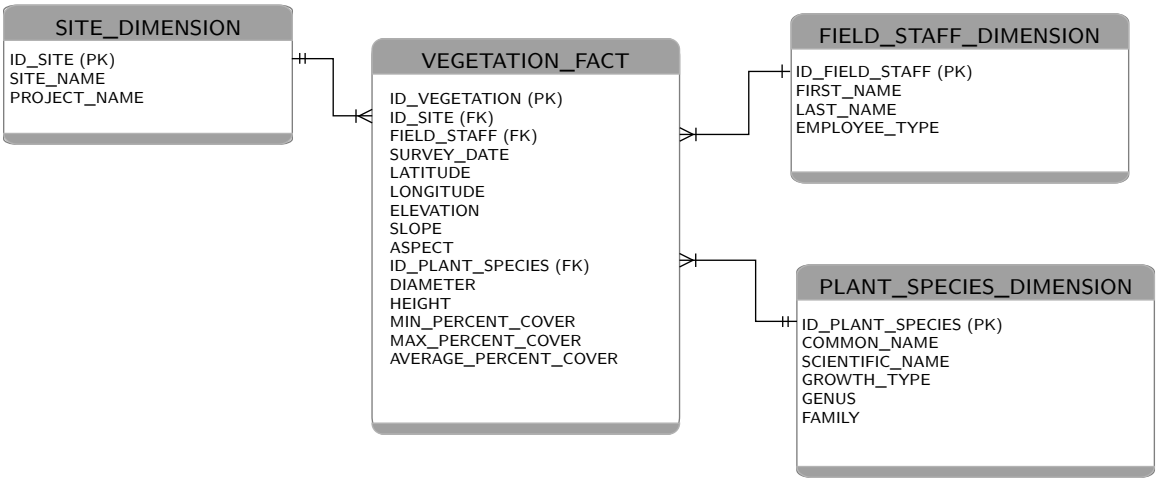


Figure 2.3: Example of a data mart designed to record vegetation survey data using the dimensional model.

The report shown previously in Table 2.1 would only reference two tables when generated from the data mart compared to the six tables referenced when generated from the transactional database.

To create a data warehouse for use in the DSS, data from multiple sources must be imported through a process of Extraction, Transformation and Loading (ETL) (Sen and Sinha, 2005). Data is first extracted from all sources. For an ecologist, this could be data collected from a number of individual research projects as well as data from other sources, such as GIS coverages. During the transformation phase data is cleaned (erroneous and missing data are corrected) and conformed to a specified standard

(El-Sappagh et al., 2011; Kimball and Caserta, 2004). In the last step, the data is loaded into a data warehouse specifically designed for the data.

Often data are also transformed in ways which are anticipated to make extraction easier. Calculated or aggregate values can be stored instead of raw measurements if this is the way the data will be used. For example, water temperature and oxygen concentration are measured in lakes and streams in order to calculate the dissolved oxygen content. It is more efficient to store this calculated value in the data warehouse rather than calculating it from the raw data each time it is required. Additionally, redundancy can be added through a process of "de–normalization" (Sen and Sinha, 2005). Whereas redundancy was purposely avoided when collecting data in a transactional data warehouse, once the data has been cleaned through ETL, values can be repeated in the data warehouse. In the vegetation survey example (Figure 2.3) latitude, longitude, elevation, slope and aspect are repeated for each plant recorded in the fact table.

Both recording context and bioacoustic data can be stored in the data warehouse. In the latter case, acoustic files must undergo a transformation so that methods of KDD can be applied.

Representation of Bioacoustic Files

Techniques employed to identify patterns (e.g. bird songs) from bioacoustic recordings can be divided into two main categories: 1) those which extract and analyze bioacoustic features and 2) those which treat a recording as a time–series set of data. Features which can be extracted from bioacoustic recordings are numerous (Table 2.2) but because most diagnostic acoustic features differ between bird species (Fagerlund, 2004), it is difficult to know which features should be extracted for a DSS which can support the identification of all species.

Table 2.2: A sample of bioacoustic features extracted for species identification from three different authors: Kirschel et al. (2009), Obrist et al. (2010) and Fagerlund (2004) (Marked in the table heading as K, O and F respectively).

Bioacoustic Feature	**K**	**O**	**F**
Rate of main part of song	✓		
Rate 1st half of main song	✓		
Rate 2nd half of main song	✓		
Number of notes	✓		
Duration start of 1st to start of 2nd note	✓		
Duration main song	✓		
Rate halves ratio	✓		
Note 2 – middle note peak frequency	✓		
Last note – middle note peak frequency	✓		
Note 2 – note 5 peak frequency	✓		
Last note – 4th last note peak frequency	✓		
Middle note peak frequency	✓		
Main song peak frequency	✓		
Minimum note peak frequency	✓		
Maximum note peak frequency	✓		
Max – min note peak frequency	✓	✓	✓
Note 2 – note 1 peak frequency	✓		
Peak frequency note 1	✓		
Peak frequency note 2	✓		
Peak frequency note 3	✓		
Peak frequency note 4	✓		
Peak frequency note 5	✓		
Peak frequency 4th last note	✓		
Peak frequency 3rd last note	✓		
Peak frequency 2nd last note	✓		
Peak frequency last note	✓		
Frequency of peak energy		✓	
Time of peak amplitude		✓	
Duration (of pulse made by bat)		✓	
Spectral Centroid			✓
Signal Bandwidth			✓
Spectral Rolloff Frequency			✓
Delta Spectrum Magnitude			✓
Spectral Flatness			✓
Zero–crossing Rate			✓
Short Time Signal Energy			✓
Modulation Spectrum			✓
Cepstral Coefficients			✓
Signal Energy Distribution in Time			✓

Alternatively, many time series classification techniques are applicable. Bagnall et al. (2017) tested 18 time series classification techniques for six different scenarios of time series. Two general approaches applicable to finding patterns within bird vocalizations are "phase independent shapelets" and "dictionary based classifiers." The first approach classifies shapelet (a short patterns that define a class) which can occur in any position within the time series. The second approach classifies events where frequency of pattern repetition is diagnostic. Both techniques could be useful when applied to the classes of bird vocalization to which they are best suited. For example, classifiers based on phase independent shapelets may be able to distinguish between House Finches (*Capodacus mexicanus*) and Purple Finches (*Capodacus purpureus*) which both have unstructured warbling songs but the House Finches includes a diagnostic burry "zreeee" note. Dictionary based classifiers may be useful for differentiating trilling birds, such as Dark–eyed Juncos (*Junco hyemalis*) and Chipping Sparrows (*Spizella passerina*) which can be distinguished by the speed each trills (the Dark–eyed Junco being faster).

Searching for similar patterns in large time series data is inefficient and several methods of transformation have been developed. One such technique, the Piecewise Aggregate Approximation (PAA), was introduced by Keogh et al. (2000) and independently by Yi and Faloutsos (2000). PAA compared favourably to other signal–reduction techniques such as Singular Value Decomposition, the Discrete Fourier Transform and the Discrete Wavelets Transform (Keogh et al., 2000) and has the additional advantage of equalizing signals that differ only in intensity (i.e. loudness) through signal normalization and smoothing intra–signal variation caused by extraneous noise (Kasten and McKinley, 2007).

Furthermore, the PAA can be converted to a symbols series which is a lower bounded approximation of the Euclidean distance of the original time series (Lin et al., 2002) through the process of Symbolic Aggregate approXimation (SAX) Lin et al. (2003). The SAX representation can then be analyzed with text and bioinformatics algorithms such as those tested by Bagnall et al. (2017) for phase independent shapelets and dictionary based classifiers. These include: Fast Shapelets (Rakthanmanon and Keogh, 2013), Shapelet Transform (Bostrom and Bagnall, 2015; Hills et al., 2014), Bag of Patterns (Lin et al., 2012), Symbolic Aggregate Approximation–Vector Space Model (Senin and Malinchik, 2013), Dynamic Time Warping with a feature generation scheme (Kate, 2016) and the Collection of Transformation Ensembles (Bagnall et al., 2015).

The PAA/SAX representation is a time and amplitude reduction of a bioacoustic file. Specifically, to reduce the size of the time dimension using PAA, a series is first divided

into segments of equal duration and an average value is calculated from each. Figure 2.4a illustrates a short segment of an acoustic recording of a single Yellow Rail (*Coturnicops noveboracensis* (Gmelin, JF, 1789)) "click". The format of this file is a Waveform Audio File Format (abbreviated WAVE or WAV) which represents analogue sound as amplitude over time.

Two steps are executed to apply PAA:

1. a time series Q is z–normalized to Q' by the formula (Kasten et al., 2007):

$$\forall i q' = \frac{(q_i - \mu)}{\sigma} \tag{2.1}$$

   where q is one of the n elements of Q, μ and σ are the mean and standard deviation of of all values of q, and q' is an element of Q'.

2. Q' is divided into equal–sized segments of size w (also known as the PAA size) where $w \leq n$. The mean of the q' values within each of these sub-sequences is then computed. Figure 2.4b shows how the same signal is still apparent once reduced tenfold through PAA.

To reduce the size of the amplitude dimension using SAX, the distribution of all amplitude values are plotted and divided into bins of equal frequency as seen along the left side of the graphs in Figure 2.4b and c. The number of bins is specified as the SAX alphabet size and values between 4 and 8 have been found to be effective (Kasten et al., 2012; Lin et al., 2003). Each amplitude bin is assigned a symbol which is assigned to each PAA segment occurring within that bin as shown in Figure 2.4c. The resulting PAA/SAX reduction for the yellow rail "click" is thus: bbbbaaaaaaaaaaaaaacaadaacaabbcbcccc. This pattern of characters is searched and analyzed by data mining algorithms developed for text.

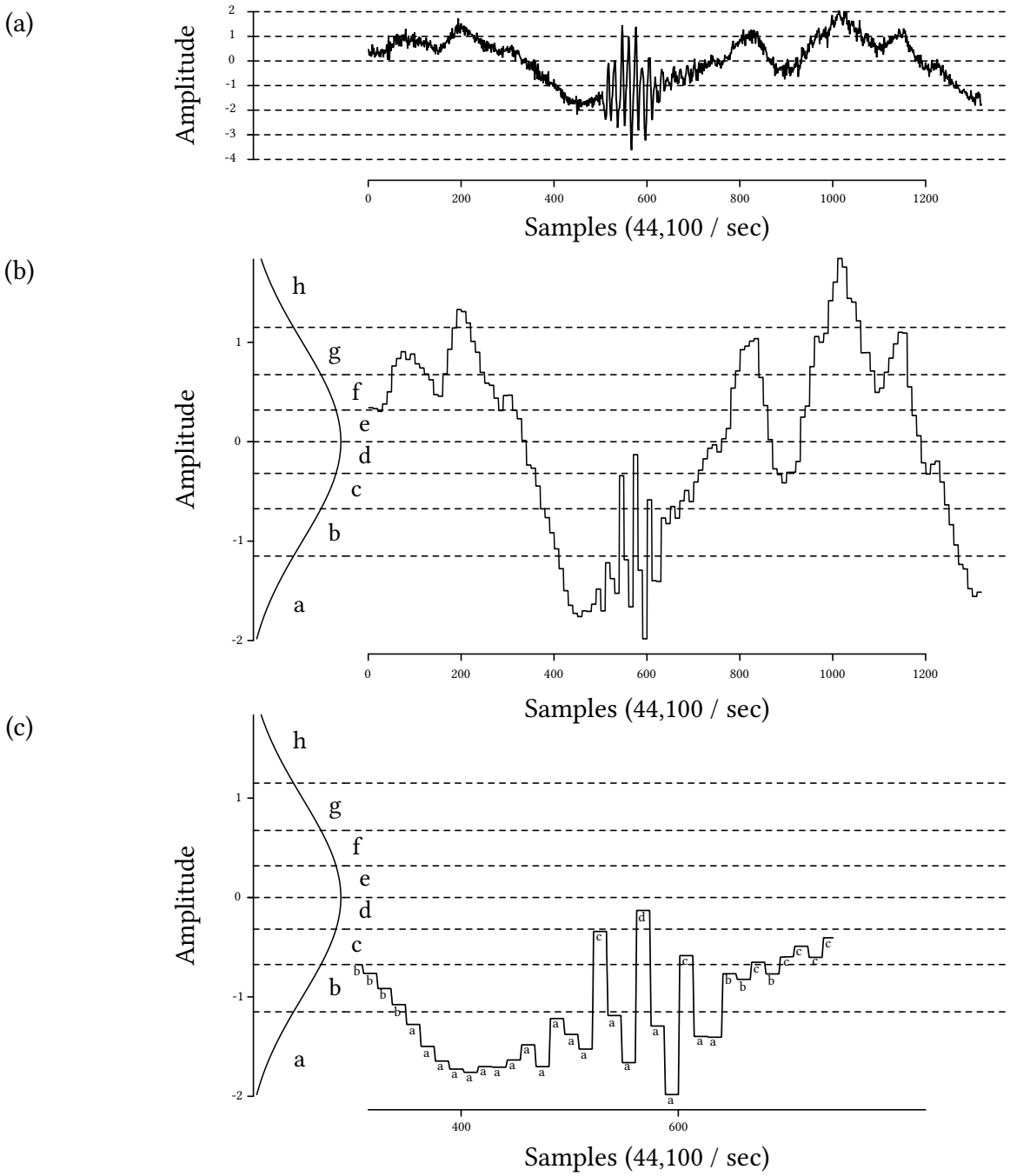


Figure 2.4: Conversion of an acoustic file (.wav format) to a symbolic representation showing how a single yellow rail “click” (a) is first reduced to a Piecewise Aggregate Approximation (PAA) (b) and assigned a symbol Symbolic Aggregate approXimation (SAX) (c) resulting in the sequence bbbbaaaaaaaaaaaaacaadaacaabbcbcccc.

Representation of Ecological Data

As stated previously, the DSS's function is to compare patterns of bioacoustic and ecological context. How ecosystem measures are included within the DSS needs to be informed by an understanding of the uncertainty inherent in ecological data which arises from unknown and complex relationships between ecosystem components (Friederichs, 1958; Beeby and Brennan, 2007) and from challenges in characterizing elements of an ecosystem (Marchini et al., 2009; Regan et al., 2012). If ignored, these uncertainties can lead to a falsely accurate representation of nature. Instead researchers should be aware of the sources of uncertainty so they can employ compensating strategies when collecting, storing and interpreting data. Knowing these limitations is crucial to meaningfully representing ecosystems in data repositories.

The causes of imprecision identified by Marchini et al. (2009) and Regan et al. (2012) can be considered within the framework of five Uncertainty Classes (UC) encountered in the stages of an ecological study (Figure 2.5) (Pouw and Kwiatkowska, 2013). These are the complexity of natural systems (UC I), the difficulty in characterizing components (UC II), the necessary assumptions required to formulate explanations (UC III), the simplifications required to represent the discovered relationships in a model (UC IV) and the challenges inherent in communication of ideas (UC V). At each of these steps, measures can be made to accommodate the expected uncertainty.

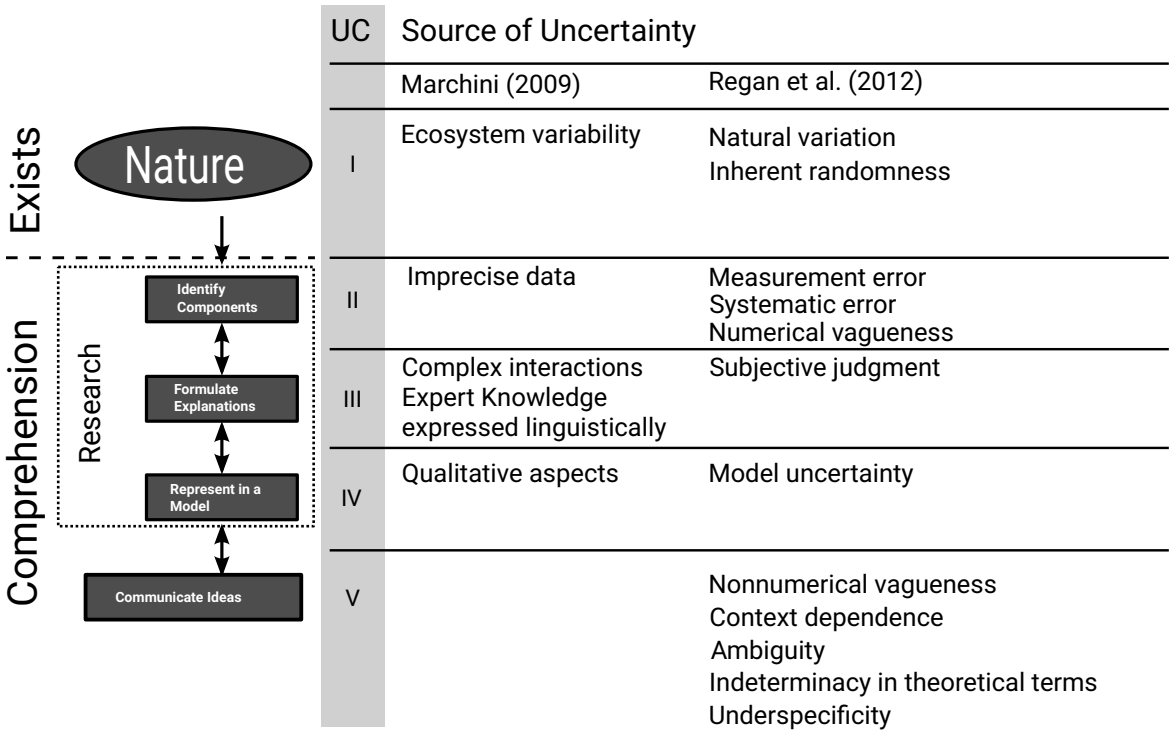


Figure 2.5: A Framework of Uncertainty within ecological studies. Arrows indicate knowledge flow. Roman numerals indicate the Uncertainty Classes (UC) discussed in the text.

As with any ecological study, collecting recording context data begins with the selection of relevant natural components (UCI) followed by a decision of how best to characterize these components (UCII). Once these decisions are made, an appraisal of how uncertainty arises and how it is dealt with must be made. For example, when the height of a tree is measured a researcher needs to understand both the limitations of the measuring technique (e.g. measurement error and numerical vagueness, from Figure 2.5) as well as the meaning of tree height as an ecosystem component. These two considerations pose related problems.

In the first case, it is difficult and tedious to measure tree height, so the value is often estimated relative to a tree that was measured by trigonometry. But, most data management systems assume that recorded values are "crisp" so an estimated height of $10\,\mathrm{m}$ is recorded as precisely $10\,\mathrm{m}$.

In the second case, what does the height value mean as a attribute of the ecosystem? Is this tree "tall" or "short"? Meaning can be assigned by classification in a system which assumed that all members of a class have share a similar ecological function. For

example, trees can be grouped into ordinal height class as described in the B.C. Forest Inventory Statistics (Table 2.3) (British Columbia Ministry of Forests, 1995).

Table 2.3: Tree height classes used by for B.C. Forest Inventory Statistics in the Cariboo Forest Region (British Columbia Ministry of Forests, 1995).

Height Class	**Height Range (meters)**
1	0 – 10.4
2	10.5 – 19.4
3	19.5 – 28.4
4	28.5 – 37.4
5	37.5 – 46.4
6	46.5 – 55.4
7	55.5 – 64.4
8	64.5+

But there are two problems which arise from using uncertain values with crisply–defined, mutually exclusive categories. Firstly, it is assumed that the component has been measured accurately and secondly, that the small change in value near a class boundary reflects a real difference in the natural world. A tree approximately $10\,\mathrm{m}$ tall can be assigned to Class 1 or Class 2, either due to a small real difference in height (here $10\,\mathrm{cm}$) or on the semi–arbitrary way its height was estimated. This characteristic of the classification system can be visualized in Figure 2.6a which shows that a tree is included completely within a category regardless of how close its height is the threshold of an adjacent category.

To accommodate imprecise data, “soft computing” techniques have been devised, such as fuzzy logic and probabilistic reasoning. Fuzzy logic was developed to represent entities which cannot be definitively categorized within a single Boolean set but can more meaningfully be considered a member of multiple sets to different degrees (Zadeh, 1965). A Fuzzy Inference System (FIS) is the practical application of the theory of fuzzy logic, composed of fuzzy membership functions and fuzzy rules (Gutiérrez-Estrada et al., 2013). Each of these components will be described using examples. The degree of membership in multiple sets will be illustrated with B.C. Forest Inventory tree heights, expression of expert knowledge using fuzzy rules will be illustrated with an assessment

how wind–like a bioacoustic recording is, and lastly a method of defining fuzzy sets from a set of data will be shown using river flow data.

Entities can be members of multiple categories because fuzzy membership functions are overlapping. Fuzzy memberships for the B.C. Forest Inventory categories could be defined as in Figure 2.6b which shows that a tree near the height threshold of two categories, has memberships in both categories.

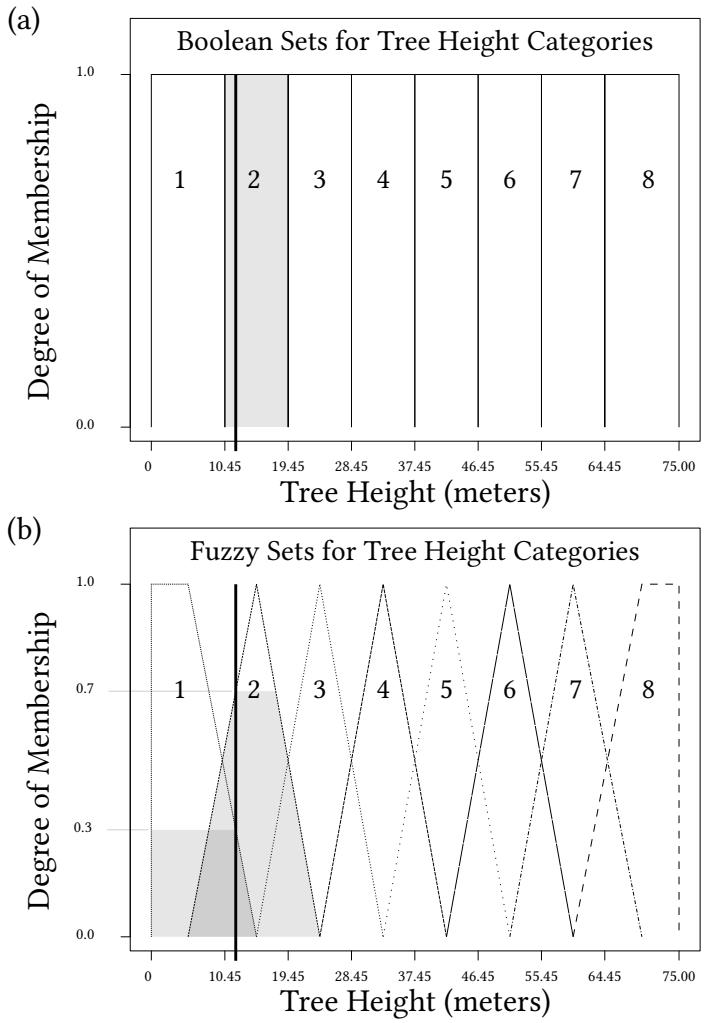


Figure 2.6: Boolean and fuzzy membership sets for the eight B.C. Forest Inventory Statistics Tree Height Classes. The bold vertical line indicates a tree with a height of $12\,\mathrm{m}$ which is included entirely within class 2 under the Boolean system and has memberships of 0.3 in class 1 and 0.7 in class 2 using the fuzzy system.

While using fuzzy sets alone can be effective to add nuanced meaning to the representation of ecosystem components, these values can be combined with expert

knowledge to interpret the significance of the data. Expert knowledge takes the form of statements using the IF, AND and OR logical operators to compile what are called Fuzzy Rules.

An example is presented of a Fuzzy Inference System for the assessment of how wind–like an acoustic event is. The fuzzy rules created are based on the work by (Towsey and Planitz, 2011) who showed that intensity and entropy features of a recording can be used to separate acoustic events caused by organisms (biophonic) from sounds caused by other natural processes (geophonic). The FIS requires eight acoustic features:

1. The maximum intensity value for any frequency below $500\,\mathrm{Hz}$: $\max\left(I_{<500\,\mathrm{Hz}}\right)$
2. The minimum intensity value for any frequency above the maximum identified in step 1: $\min\left(I_{>500\,\mathrm{Hz}}\right)$
3. The frequency of the maximum mean intensity: $f_{max(I)}$
4. The frequency of the minimum mean intensity: $f_{min(I)}$
5. The minimum entropy value for any frequency located below $500\,\mathrm{Hz}$: $\min\left(S_{<500\,\mathrm{Hz}}\right)$
6. The maximum entropy value for any above the location of the minimum identified in step 5: $\max\left(S_{>500\,\mathrm{Hz}}\right)$
7. The frequency or maximum entropy: $f_{max(S)}$
8. The frequency of the minimum entropy: $f_{min(S)}$

Table 2.4 shows the acoustic features which are greater for wind than for biophonic events. To create a partial FIS to determine how wind–like an acoustic event is based on the first two parameters of Table 2.4, fuzzy membership functions are created for the differences in frequency and intensity, each with two overlapping sets for a “small” and a “large” difference. An output set of fuzzy memberships is created for “wind–like” with sets for “not”, “somewhat” and “very” wind–like (the Input Membership Functions in Figure 2.7).

Next, each of the acoustic features listed in Table 2.4 is expressed in an FIS as a fuzzy rule. The four possible fuzzy rules for the first two acoustic parameters, called here the

"intensity difference" and "frequency difference" are shown in Table 2.5. The possible conditions allowed for each input parameter are "small" and "large". The predictions of the FIS are that the sound is "not," "somewhat" or "very" wind–like. For the purpose of this example we assume that if both features are "small" then the sound is "not" wind-like, if both features are "large" the sound is "very" wind–like and if the magnitude of one acoustic features are "small" and the other "large" the sound is "somewhat" wind–like.

Table 2.4: Acoustic features for the identification of acoustic events caused by wind where I is intensity , f is frequency and S is entropy of an acoustic signal. From Towsey et al. (2012).

Parameter No.	**Definition of Feature**	**Magnitude for Wind vs. Non–Wind events**
1	$\max\left(I_{<500\,\mathrm{Hz}}\right) - \min\left(I_{>500\,\mathrm{Hz}}\right)$	greater
2	$f_{max(I)} - f_{min(I)}$	greater
3	$\min\left(S_{<500\,\mathrm{Hz}}\right) - \max\left(S_{>500\,\mathrm{Hz}}\right)$	greater
4	$f_{max(S)} - f_{min(S)}$	greater

Table 2.5: Fuzzy rules for the detection of wind using parameters no. 1 and 2 of the four parameters identified by Towsey et al. (2012).

Rule No.	**IF intensity difference is...**	**(Logical Operator)**	**IF frequency difference is...**	**THEN the sound is wind–like...**
1	small	AND	small	not
2	small	AND	large	somewhat
3	large	AND	small	somewhat
4	large	AND	large	very

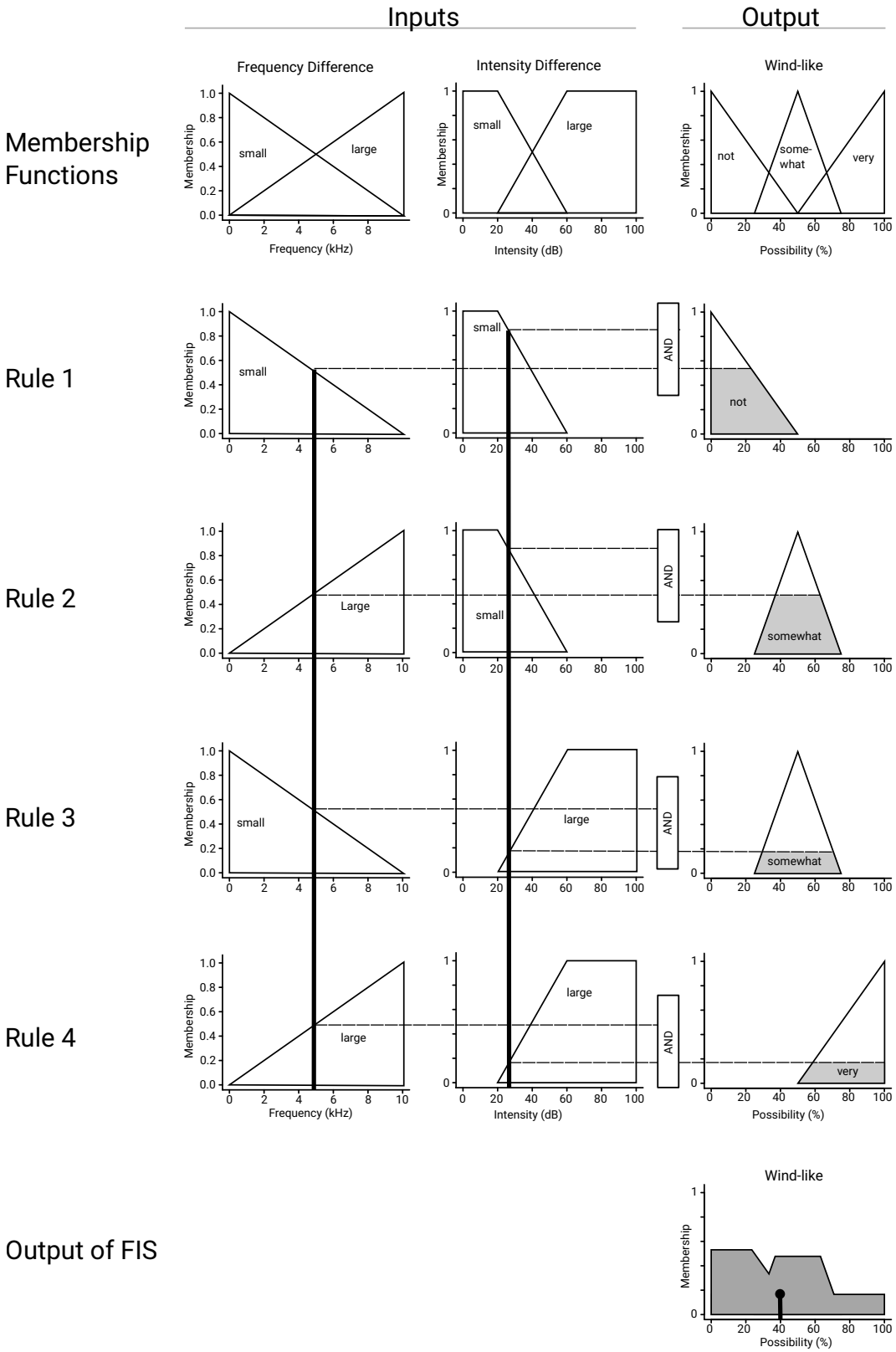


Figure 2.7: A fuzzy inference system to categorize acoustic signals as wind–sound. Two parameters of the four used by Towsey et al. (2012) are shown. An example is shown with input values: frequency difference $= 4.691$ kHz, intensity difference $= 26.2$ dB. Output: wind–like $= 39.4\%$.

For each rule, the degree to which each corresponding input membership functions represents the output case is evaluated. Rows 2 to 5 of Figure 2.7 illustrates each rule of the FIS being applied. For example, rule 1 shows how the degree to which the frequency difference and intensity difference are “small”. The AND operator selects the lowest degree of membership of the two and applies this as the membership of the “not” wind–like function (shaded area). When all output memberships are determined a single output value for the FIS is derived by calculating the geometric centre of the polygon created by all overlapping output memberships (Output of FIS in Figure 2.7). In this example, a frequency difference of 4.691 kHz and an intensity difference of 26.2 dB is found to be 39.4% wind–like.

A full FIS using all four parameters from Table 2.4 would increase the number of rules to 16 (2^4 for four rules each with two membership functions) and would require five fuzzy sets in the output based on the possible combination of large and small memberships.

The number, shape and boundaries of fuzzy membership sets are not defined by the theory but are selected by the practitioner. Two or more sets can be defined, they can be triangular, trapezoidal or Gaussian and can be defined for any range within the universe of domain values. Where expert knowledge exists, membership functions can be created to reflect this. Another option is to be guided by the distribution of measured values.

Using the technique described by Suh (2012), membership functions can be created from the data to be classified if the distribution is close to a normal distribution. An example of this technique is shown using water flow, a common ecological measure when studying rivers. Figure 2.8a shows a histogram of flow measurements for an example data set (Durbin and Koopman, 2012). The distribution is approximately normal as can be seen from a Q–Q plot (Figure 2.8b).

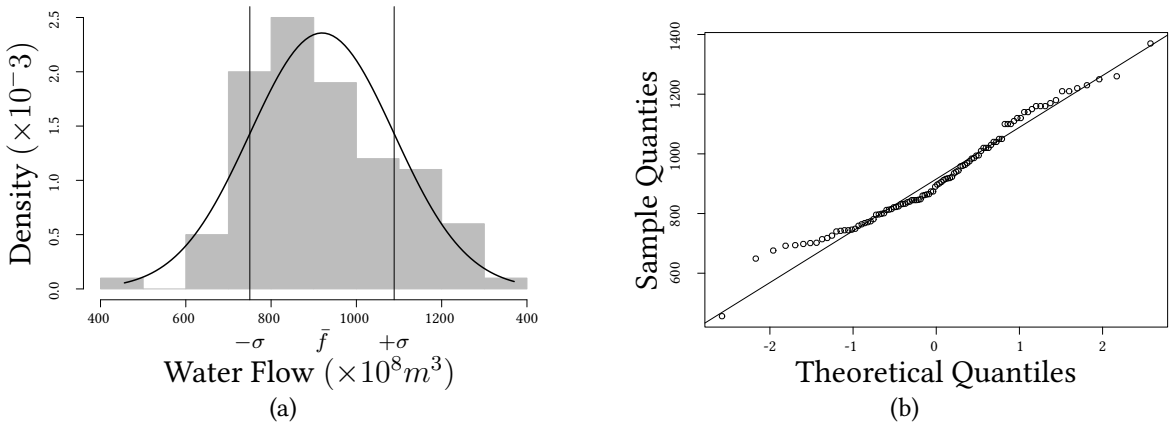


Figure 2.8: Distribution of water flow data used to derive fuzzy membership sets. The Q–Q plot shows that the distribution is approximately normal.

The mean, minimum, maximum and standard deviation ($\bar{f},\ f_{min},\ f_{max}$ and σ) are calculated and are then used to calculate the mean of two subsets of values greater–than and less–than one standard deviation away from the mean ($\bar{f}_{low}$ and $\bar{f}_{high}$) using Equation 2.2 and Equation 2.3 (Table 2.6).

$$\bar{f}_{low} = \frac{1}{n}\sum_{i=1}^{n}\left\{f : f < \bar{f} - \sigma\right\} \tag{2.2}$$

$$\bar{f}_{high} = \frac{1}{n}\sum_{i=1}^{n}\left\{f : f > \bar{f} + \sigma\right\} \tag{2.3}$$

Table 2.6: Values used to define fuzzy sets based on the characterization of flow measurements.

Value	Flow $(\times 10^8 m^3)$	Description
$\bar{f}_{min}$	456	Minimum flow value
$\bar{f}_{low}$	699	Mean of value less than one standard deviation below the mean
$\bar{f} - \sigma$	750	One standard deviation below the mean
$\bar{f}$	919	Mean flow value for all data
$\bar{f} + \sigma$	1,089	One standard deviation above the mean
$\bar{f}_{high}$	1,174	Mean of values one standard deviation above the mean
$\bar{f}_{max}$	1,370	Maximum flow value

The five values of $f$are then used to define the horizontal coordinates of two trapezoidal and one triangular membership functions while the vertical coordinates are given a value of either zero or one (Figure 2.9).

Calculating the membership of a flow value in each of the three fuzzy sets can be done graphically, as was shown in Figure 2.9, or using Formulas Equation 2.4, Equation 2.5 and Equation 2.6.

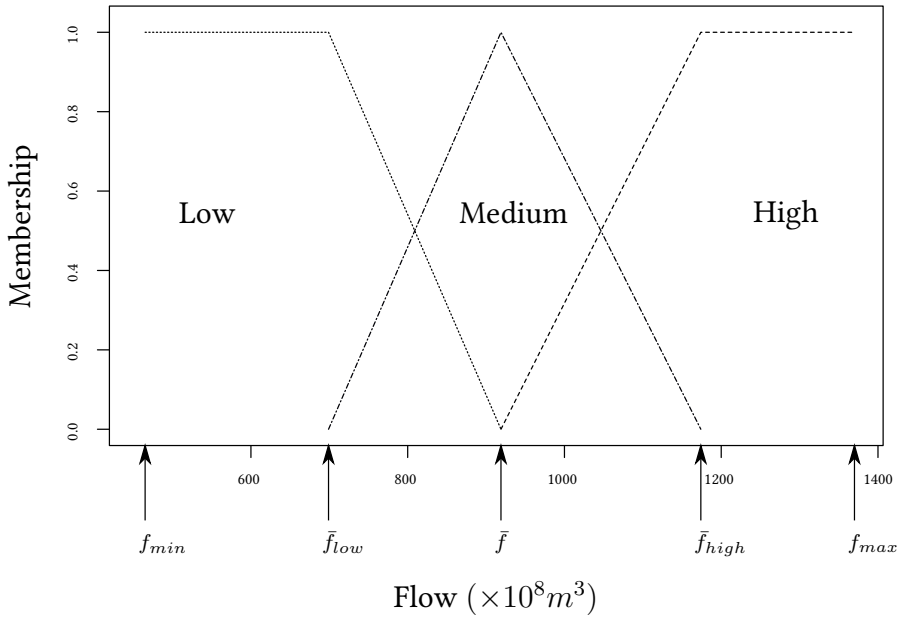


Figure 2.9: Derivation of membership functions from normalized distribution.

$$\mu_{low}(f) = \begin{cases} 1 & if \quad f_{min} \leq f \leq \bar{f}_{low} \\ \frac{\bar{f}-f}{\bar{f}-\bar{f}_{low}} & if \quad \bar{f}_{low} \leq f < \bar{f} \\ 0 & if \quad f > \bar{f} \end{cases} \tag{2.4}$$

$$\mu_{medium}(f) = \begin{cases} 0 & if \quad f < \bar{f}_{low} \\ \frac{f-\bar{f}_{low}}{\bar{f}-\bar{f}_{low}} & if \quad \bar{f}_{low} \leq f \leq \bar{f} \\ \frac{\bar{f}_{high}-f}{\bar{f}_{high}-\bar{f}} & if \quad \bar{f} < f < \bar{f}_{high} \\ 0 & if \quad f \geq \bar{f}_{high} \end{cases} \tag{2.5}$$

$$\mu_{high}(f) = \begin{cases} 0 & if \quad f_{min} \leq f \leq \bar{f} \\ \frac{f-\bar{f}}{\bar{f}_{high}-\bar{f}} & if \quad \bar{f} \leq f < \bar{f}_{high} \\ 1 & if \quad f \geq f_{high} \end{cases} \tag{2.6}$$

Discussion

The advantages of data warehouses have not been realized by many ecologists who typically store their data in spreadsheets or transactional databases. Often their data is housed in a number of separate files, each one created specifically for individual research projects, making it difficult to consider the data as a whole. By integrating these data sources within a single data warehouse, ecologists have the opportunity to search through all their data to discover “valid, novel, potentially useful and ultimately understandable patterns in data” through Knowledge Discovery in Databases (Fayyad, 1997). A specific application of a data warehouse has been presented here; as a component of a Decision Support System to aid human listeners to process bioacoustic recordings. Any ecologist can benefit from compiling their data to into a data warehouse to discover relationships which can inspire future research. However, the process of moving to a data warehouse may pose challenges to an ecologist.

Ecological data is arguably more diverse than the financial transactions which data warehouses were originally designed for. Ecosystems are complex and often not fully understood, ecological parameters can be difficult to define and logistically challenging to measure. Consequently, representing ecological data in any data management system requires careful evaluation of the sources of uncertainty and consideration of methods which could best represent the ecosystem component. Fuzzy logic is one example of a soft-computing technique that has been applied to represent instances of uncertainty but is not yet commonly used in this field.

Ecologists have seen a dramatic expansion in the volume of data they must manage from the increased use of computer–based data loggers which has necessitated the need for automated or semi–automated data processing. For example, Autonomous Recording Units have drastically increased the amount of bioacoustic data which must be analyzed. Data warehouses are well suited to store this volume of data but it should be stored in a form which data mining algorithms can process. One simple method is to treat the recording as a time series and to use aggregation techniques to reduce the signal to a sequence of symbols through PAA/SAX. The resulting format is well suited to many data mining routines.

Even with the challenges inherent in ecological data, researchers with multiple data sets can benefit from developing a data warehouse because it can extend the use of their data. In the next chapter the process of creating a data warehouse from existing bioacoustic data will be described.

Chapter 3

Pre-processing and Organization of Ecological Data to Facilitate Knowledge Discovery

Collecting ecological data can be difficult because of logistics, difficulty in measuring ecological components and human error introduced by field technicians. The ability to get to sites can be challenging, especially to remote locations and habitats that are difficult to access or traverse. Once at a site, measuring habitat components can be unrealistic where conditions are unsafe, components are difficult to measure or because real–world conditions do not match predefined categories. Additionally, field measurements are often done by students who frequently are required to make estimates based on judgment despite their inexperience in the applicable protocol. Consequently, ecological data can be incomplete, contain errors and imperfectly represent the habitat they were meant to quantify.

In contrast, business data can seem simple to gather and manage. Business transactions are often logged in real–time, are definitive and don't rely on human judgment so the application of techniques designed for business data, such as data warehouses, to relatively "messy" ecological data can pose unique challenges. In this chapter a data warehouse is created for an existing ecological data set of bioacoustic research. The steps chosen to prepare the data and design the data warehouse as a component of the proposed DSS are outlined in the Methods section. The Results section relates the specific problems encountered with the data as the steps in the Method were applied. Recommendations are made in the Discussion for moving ecological data into a data warehouse based on the experience gained in this research.

Methods

To systematically create a data warehouse for bioacoustic data, the steps of Knowledge Discovery in Databases procedure (KDD) (Azevedo and Santos, 2008) were followed. Table 3.1 outlines the steps of this procedure. KDD step 1 involves the collection of the field and reference data required. KDD step 2 is a necessary step to move "clean" data from the Input to the Codifying and Organization steps of the DSS, described in Chapter 2. Techniques employed in KDD step 3 simplify data management and add knowledge to the data. This is accomplished through averaging values, collapsing hierarchical data structure and applying specialized techniques such as fuzzy logic where appropriate.

These first three steps were executed prior to the development of the data warehouse and will be described in the remainder of this chapter. The applications of the remaining four steps describe the Pattern Matching, Reports and User Input sections of the proposed DSS (Figure 2.1) and represent the next–steps in the DSS.-

Table 3.1: Steps of the Knowledge Discovery in Databases (KDD) procedure.

Steps
1. Selection of relevant data though an understanding of data and the goals of the KDD
2. Processing of data to handle missing or erroneous values
3. Reducing the dimensionality (number of attributes) in the data by aggregation
4. Selecting data mining techniques (clustering, summation, modelling, classification and change detection)
5. Searching for novel and non-trivial patterns
6. Interpreting discovered patterns
7. Incorporating the knowledge in a models

Knowledge Discovery in Databases Step 1 - Selection of Relevant Data

The data used in this thesis were made available by Dr. Erin Bayne of the University of Alberta, who maintains the Ecological Monitoring Committee of the Lower Athabasca region (EMCLA) database. The database was comprised of data and metadata pertaining to the deployment of ARUs, the creation of bioacoustic recordings and the identification

of species from those recordings. Data were collected for a number of projects focused either on studying individual species of interest or to compare the biodiversity of ecosystems. The spatial extent of the data base covered Alberta and included parts of British Columbia and the Northwest Territories. To reduce the size of the data set for testing, only ARU deployments within the Lower Athabasca region were selected (Figure 3.1).

ARUs were usually used in groupings to study an effect such as a man-made disturbance. For example, Figure 3.2 illustrates how ARUs could be used to study the effects of roads on bird populations. One ARU is deployed near the road where road effects would be strongest, two more ARUs are placed at increasing distances from the road where the road effects would be weaker. These three ARUs would be considered the experimental treatment group. Additional sets of ARUs would be deployed far enough away from roads to function as experimental controls.

An explanation of the nomenclature used in the bioacoustic data and a description of each component of the EMCLA data follows:

The location where each ARU is installed is known as a *Site* and each group of ARUs is called a *Station*. All treatment and control Stations are combined to make a *Project*.

Each ARU would be set to capture an acoustic *Recording* at a preset interval for a preset duration. From *Acoustic Events* on these recordings, a bird species would be identified and recorded as a *Detection*.

Because an ARU may be deployed at a Site repeatedly within a year or over subsequent years, each occurrence of an ARU installation was called a *Deployment*.

Figure 3.1: (a) The Lower Athabasca region of Alberta (shaded) within the province of Alberta, showing locations of ARU deployments (dots).
(b) Lower Athabasca region of Alberta showing locations of ARU deployments (dots).

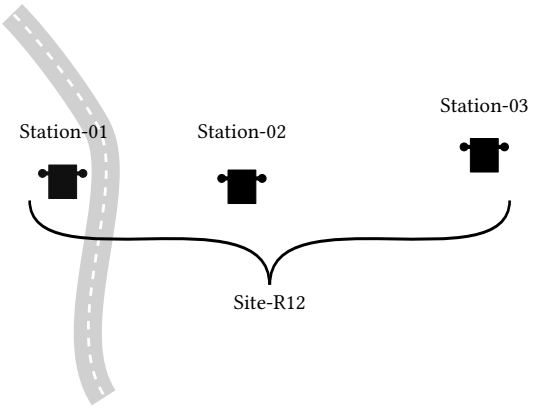


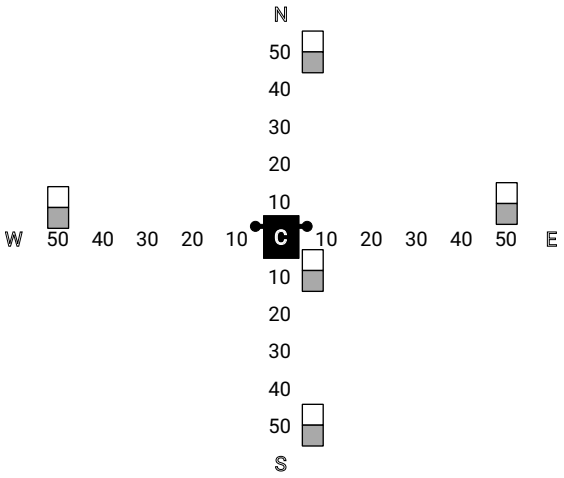
Figure 3.2: An example of three ARUs used to study the effects of roads on bird distribution. The location where each ARU is placed is called a *Station*, the group of ARUs is called a *Site*. Several Sites would be sampled as part of a *Project*. The ARUs are placed so that Stations 01 and 02 measure effects close and far from the road while Station 03 is the control, placed outside the road effects.

The data extracted from the EMCLA database represent the application of KDD step 1, the selection of relevant data. They provided the bioacoustic, temporal–context and spatial–context data for the Decision Support System (Figure 2.1).

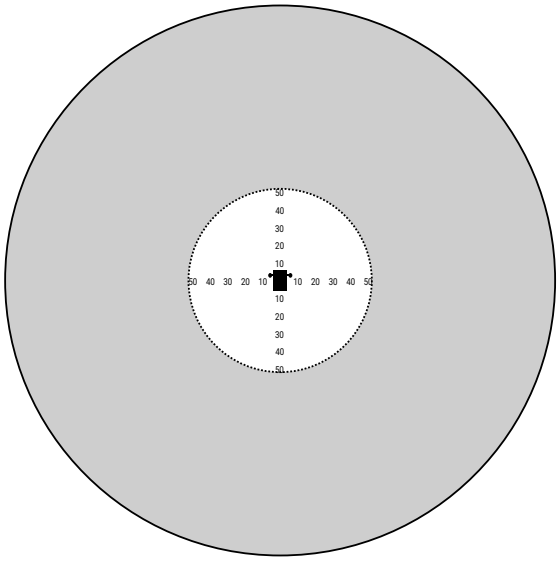
Specifically, acoustic file names and species identified from these recordings represent bioacoustic information. The bioacoustic recordings themselves were not stored in the EMCLA database but were downloaded from a networked repository maintained by Dr. Bayne. These ten-minute stereo recordings were converted from the proprietary "WAC" (.wac) compressed format to a wave (.wav) format using Wildlife Acoustics's *wac2wav* conversion software. Files had been recorded at a bitrate of 44,100 kHz. A list of species names and codes used by the listeners was also downloaded. This list included the taxonomic levels from order to species, common names in English and the standard 4-letter code used by the American Ornithological Union (North American Classification Committee, 2014) as well as true/false fields for the species occurrence in Alberta and its vocalizations characteristics as determined by University of Alberta staff.

The temporal components of recording context were available from the recording metadata. Time-of-season and time-of-day data were provided by the date and time when a recording was made. Additionally, the period between the ARU deployment and retrieval date was also available for time-of-season knowledge. Time was recorded in civil time which in Alberta is Mountain Standard or Mountain Daylight time.

The spatial components of recording context was provided by the geographic coordinates for all ARU deployments. For a subset of deployments, additional spatial information was available in the form of field assessments of water depth, wetland type and horizontal cover. Water depth measurements were made at and around an ARU deployment site at 21 points, $10\,\mathrm{meters}$apart in the cardinal directions (Figure 3.3a). Wetland type was judged according to the Ducks Unlimited Enhanced Wetland Classification (DUEWC)(Ducks Unlimited, 2015) system (Table 3.2). Observed wetland classes were recorded in two circular regions, within $50\,\mathrm{meters}$ around the ARU deployment and between $50\,\mathrm{m}$ and $150\,\mathrm{m}$ from the ARU deployment (Figure 3.3b). Horizontal cover was estimated using a $1\,\mathrm{m} \times 2\,\mathrm{m}$cover board, divided into two equal squares (Figure 3.4). From a distance of $50\,\mathrm{m}$ the percentage of each square obscured by intervening vegetation was estimated. This process was repeated at five points at and around the ARU deployment (Figure 3.3a).



(a) Pattern at which water depths were measured around an ARU. Numbers indicate the distance in meters from the central point "C", where the ARU is placed.
Horizontal cover was measured from point "C" to each 50 meter point and from the northern 50 meter point back to point "C", as marked by the grey and white rectangles.



(b) Areas around ARU deployment stations which were classified by the DUEWC system. The white circle in the centre shows 50 meter area, the grey region shows the area between 50 meters and 150 meters radius.

Figure 3.3: Habitat assessment at ARU deployment locations.

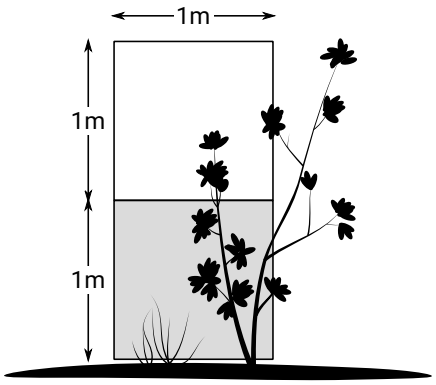


Figure 3.4: Sketch of a horizontal cover board in use. An estimate is made of how much each of the top and bottom squares are obscured by vegetation. In this example the top square is approximately 10% obscured and the bottom square it approximately 30% obscured.

Table 3.2: Standard Ducks Unlimited Enhanced Wetland Classification (DUEWC) codes provided to field technicians.

DUEWC Description	DUEWC Code
Conifer Upland	UCN
Deciduous Upland	UDC
Mixedwood Upland	UMX
Other Upland	UOT
Treed Bog	BTR
Shrubby Bog	BSH
Open Bog	BOP
Treed Poor Fen	FPT
Shrubby Poor Fen	FPS
Graminoid Poor Fen	FPG
Treed Rich Fen	FRT
Shrubby Rich Fen	FRS
Graminoid Rich Fen	FRG
Shallow/Open Water	WAT
Emergent Marsh	MEM
Meadow Marsh	MMD
Shrub Swamp	SSH
Hardwood Swamp	SHR
Mixedwood Swamp	SMX
Tamarack Swamp	STM
Conifer Swamp	SCN

ARU deployment data were downloaded for this research on 2016/07/27. EMCLA data were extracted by queries to the University of Alberta's SQLServer database via a MS Access database. Table 3.3 shows the type and number of records retrieved. Records were downloaded corresponding to 7,933 ARU deployments from which 18,741 recordings were made and 122,743 species identifications determined. A list of 435 candidate species was also downloaded. In-field habitat assessments were available for some ARU deployment records: water depth (736), wetland type (534) and horizontal cover (449). To ensure that only data which had been validated by U of A staff was considered, only the 3,533 records for deployments prior to 2014 were considered. The results of these queries were stored as comma-delimited text files which were then uploaded to tables in an Oracle database into "raw" data tables.

Table 3.3: Records extracted from the EMCLA database for use in the Decision Support System. The number of records listed are from before the data were subset by time or location.

Theme	Information Type	Use in the DSS	Number of Records
Species Codes	Order Family Genus Species English Name 4-Letter Code In Alberta (Y/N) Call Type	Bioacoustic	435
Species Identified	File Name Species Code	Bioacoustic	122,743
Recordings Made	Deployment Information File Name Date & Time of recording Start	Bioacoustic Temporal-Context	18,741
ARU Deployment	Project Identification Geographic Location Installation Date Retrieval Date	Temporal-Context Spatial-Context	7,933
Water Depth	Deployment Information Plot Location Depth Measurements	Spatial-Context	18,787
DUEWC Wetland	Deployment Information Buffer Distance Habitat Class Percent Area Estimate	Spatial-Context	2,052
Horizontal Cover	Deployment Information Plot Height Percent Obscured	Spatial-Context	5,067

Secondarily Derived Information

Because field assessments were only made for a proportion of ARU deployments, additional spatial-context attributes were derived through integration of the EMCLA data with CanVec+ habitat maps (Natural Resources Canada, 2014) and the Alberta Digital Elevation Model (DEM) (Ministry of Natrural Resources, 1997) using GRASS GIS (GRASS Development Team, 2015).

ARU deployment coordinates from the EMCLA data set were loaded into the GIS and those within the Lower Athabasca region (1,819 deployments) were extracted. The results were then exported to a temporary table in the Oracle database and used to identify deployment records within the region of interest.

The GIS was also used to create circular buffers around each deployment site. A radius of $150\,\mathrm{meters}$ was chosen to match the DUEWC habitat assessments. These buffers were used to overlay the CanVec+ Wooded Area entity in the Vegetation theme (Table 3.4), the Wetland entity in the Saturated Soil theme and the Water Body entities in the Hydrology theme. The GIS reported the area (m^2) of each habitat, these values were imported into the ORACLE database as a proportion of the total buffer area. Because layers of the CanVec+ coverages overlap, total habitat proportions could add up to more that one.

Using the *r.terraflow* tool and the Alberta DEM in the GRASS GIS, a water accumulation map was created (Figure 3.5). The area (in m^2) of terrain draining into each deployment site was then extracted from the GIS to the Oracle database.

An additional temporal-context attribute was derived by calculating time-before and time-after local sunrise as well as time-before and time-after local sunset using the *maptools* package (Bivand and Lewin-Koh, 2016) for the R statistical program (R Core Team, 2017) with the recording-time and deployment coordinates from the EMCLA database as parameters. This data was also imported into the Oracle database.

Table 3.5 summarizes all secondary information derived for inclusion into the Oracle database for the Field Data section of the DSS.

Table 3.4: Canvec+ Wooded Area entity attributes.

Codes	Code Label
0	No Data
10	Unclassified
11	Cloud
12	Shadow
50	Shrubland
51	Shrub Tall
52	Shrub Low
81	Wetland Treed
200	Forest/Tree classes
210	Coniferous Forest
211	Coniferous Dense
212	Coniferous Open
213	Coniferous Sparse
220	Deciduous Forest
221	Broadleaf Dense
222	Broadleaf Open
223	Broadleaf Sparse
230	Mixed Forest
231	Mixewood Dense
232	Mixewood Open
233	Mixewood Sparse

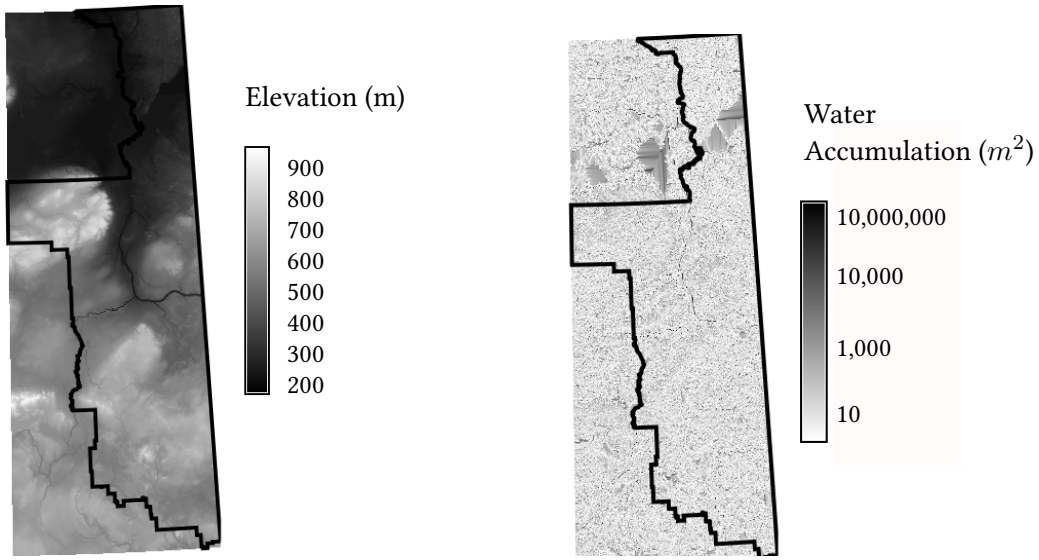


Figure 3.5: Topography and Hydrology of the Lower Athabasca Region, (a) Digital Elevation map (b) Water accumulation.

Table 3.5: Secondary information derived from other data sources and data transformations.

From EMCLA Database	From External Sources	Derived Data
Deployment Coordinates	CanVec+[a] themes: Wooded Area Saturated Soil Hydrology	Terrestrial Habitat Wetland Wooded Area
Deployment Coordinates	Water Accumulation Map (created from the Alberta Digital Elevation Model[b])	Water Accumulation
Deployment Coordinates, Recording Time	—	time before sunrise time after sunrise time before sunset time after sunset

[a] Natural Resources Canada (2014)
[b] Ministry of Natrural Resources (1997)

Knowledge Discovery in Databases Step 2 - Processing of Data to Handle Missing or Erroneous Values

The second step of the KDD procedure is to process data for missing and erroneous values. This step is critical to obtain valid results from data queries. Records of deployments, recordings and species detections were omitted from the study if data necessary to specify date, time or identity were missing. Erroneous data were identified by comparison with the domain of expected values based on the research protocols under which the data were collected (Table 3.6).

Additional fields were added to the raw data tables in the Oracle database. These fields were called IS_VALID, IN_TIMEFRAME and IN_REGION . These fields were set to 0 for all records. After analysis, valid records were assigning a values of 1. The methods used to identify and mark missing and erroneous data were specific to the type of data considered. The next sections describe these methods.

Table 3.6: Valid Ranges of Data for KDD step 2.

Data	Units	Domain	Expected Number of Values
Deployment / Retrieval Date and Time (t_d, t_{rt})	year/month/day/ hour:minute:second	$t_d, t_{rt} \in [2000/01/01, 2014/12/31]$ $t_d < t_{rt}$	1 deployment data and 1 retrieval date per deployment
Recording Date and Time (t_{rc})	year/month/day/ hour:minute:second	$t_{rc} \in [2000/01/01, 2014/12/31]$ $t_d < t_{rc} < t_{rt}$	1 per recording
Location (l)	latitude & longitude	$l_{latitude} \in [-90°, 90°]$ $l_{longitude} \in [-180°, 180°]$ $l \subseteq \{Lower\ Athabasca\ Region\}$	1 set of coordinates per ARU deployment
Species	none	Codes used by the University of Alberta	1 species per detection record
Water Depth (d)	cm	$d \in [0, 100]$ for aggregation $d \in (0, 100]$ for distribution	21 per ARU deployment
Wetland Assessment $(DUEWC)$	n/a	DUEWC categories	1 (inner buffer), 1 (outer buffer) per ARU deployment
Horizontal Cover (c)	percent	$c \in [0, 100]$	5 high & 5 low per ARU deployment
Flow Accumulation (f)	m^2	$f \geq 0$	1 per ARU deployment
Habitat Type (CanVec+ categories)(h)	none (proportion)	$h \in [0, 1]$	1 or more per ARU deployment
Sun-time(t_s)	minutes	$t_s \in [-1440, 1440]$	1 sunset & 1 sunrise time per recording

Dates

First, a range of dates was selected to omit records which had not yet been validated by University of Alberta technicians. These records were marked with a 1 value in the IN_TIMEFRAME field.

Within this subset of data, dates which were either erroneous or suspicious were sought through database queries according to the following classes:

Class 1 Missing Values (null) Deployment date or Retrieve date is NULL.

Class 2 Out-of-Range Deployment date or Retrieve date outside a reasonable threshold

Class 3 Out-of-sequence Retrieval dates proceed Deployment dates.

Class 4 Range Greater than Likely Retrieval date follows deployment date by an unlikely length of time.

Records which were identified as Class 1, Class 3 were excluded in all cases. Records identified in classes 2 and 4 were manually evaluated.

Locations

Records were selected where latitude and longitude coordinates for ARU deployment were present. These deployment records were then imported into the GIS where those which lay within the Lower Athabasca region were selected. This subset was then exported back to the Oracle database and was used to set the field IN_REGION field to 1. In this way both deployments outside of the study region and any with invalid or missing coordinates were omitted.

Species

For records where species detections had been made, the four-letter species code recorded were matched to the list of codes downloaded from the EMCLA database. These codes included the American Ornithologists' Union (AOU) species codes for birds

(North American Classification Committee, 2014) as well as codes for amphibians and mammals. Where recorded species codes did not match reference codes, an interpretation of the incorrect code was made based on the researcher's birding experience. All species codes that matched were set to IS_VALID = 1.

Water Depth

At deployment sites where water depth had been measured (with a meter stick), values between $0\,\mathrm{cm}$ to $100\,\mathrm{cm}$ were considered valid and were used in subsequent aggregation methods. For deployments where some depth values were missing or invalid, calculations were made with the remaining values.

Wetland Assessment

Habitat assessments were compared to the valid DUEWC codes (Table 3.2). Codes which were invalid were matched to appropriate DUEWC classes where possible. In other cases, all valid information about the wetland classification was retained. If no match could be made with standard categories, no wetland habitat information was applied to that record.

Horizontal Cover

Horizontal coverage estimates could range from completely un-obscured (0%) to completely obscured (100%). Values outside of this range were excluded from subsequent aggregation calculations. On sites where some values were missing or invalid, the remaining values were used to calculate aggregate values.

Knowledge Discovery in Databases Step 3 - Reducing Dimensionality

Reducing dimensionality of data (KDD Step 3) simplifies data management and adds knowledge to the data. Dimension reduction was applied to acoustic recordings, water depth, horizontal cover and wetland classification using methods of aggregation appropriate for each data type (Table 3.7).

Table 3.7: Aggregation of habitat data.

From EMCLA Database	Transformation / Aggregation
Acoustic Recordings	PAA/SAX
Water Depth	Averages & Fuzzy Memberships
Horizontal Cover	Averages
DUEWC Habitat	Proportion

Acoustic Recordings

Acoustic recordings were reduced through Piecewise Aggregate Approximation (Keogh et al., 2000) followed by Symbolic Aggregate approXimation (SAX) (Lin et al., 2003) using the *sax_by_chunking* method in the *jmotif* package (Senin, 2016) for the R statistical program. The parameters used were an alphabet size of 8 and a PAA size (w) calculated to reduce the acoustic file by a factor of 10 (Equation 3.1).

$$w = round(length(recording)/10) \tag{3.1}$$

The minute-long symbolic representations were imported into a temporary table in the database and concatenated to represent the original ten-minute recording before being stored as a Character Large Object (CLOB).

Water Depth

The 21 water depth measurements for each ARU station were summarized both as a simple mean and as a membership in three fuzzy sets (shallow, medium and deep). Triangular fuzzy sets were derived from the normalized distribution of depth values following the method outlined by Suh (2012). Because all plots without water depth were given a value of $0\,\mathrm{cm}$ (i.e. the distance above the water table was not measured) only the distribution of values greater than $0\,\mathrm{cm}$ (referred to "wet plots") were used for fuzzy set derivation, as described in the section.

Horizontal Cover

The five sets of low and high horizontal cover for each ARU station were aggregated as a mean for each height class as well as a combined mean.

Habitat Index

The two DUEWC habitat index values recorded at ARU stations were aggregated to a single value. However, since the individual estimates reflected the proximity of habitat to the site, they were retained in the database.

Implementing the Dimensional Design Process

Between KDD step 3 (Reducing Dimensinoality) and KDD step 4 (Selecting a Data Mining Techniques), it is beneficial to arrange the data in a structure optimized to extract data for data mining techniques. By following the steps of the Dimension Design Process (DDP) (Kimball et al., 2002) (Table 3.8), a collection of data marts was created to support such *ad hoc* queries.

Recall that within a data mart, quantitative attributes are stored in a single fact table while qualitative attributes (used for filtering and grouping) are stored in related dimension tables (See Chapter 1). The collection of data marts comprise a data warehouse.

Table 3.8: Steps of the Dimension Design Process (DDP).

Steps
1. Select the research focus
2. Declare the grain
3. Identify the dimensions
4. Identify the facts

The development of the bioacoustic Decision Support System was adopted as the focus of the data mart design process (DDP step-1). The number of data marts and the granularity of each was chosen to reflect the major research activities (DDP step 2). Data

related to the dimension and fact tables for each was identified, appropriate tables were built in the Oracle database and data was copied from the raw data tables into these new tables (DDP step 3 and 4). The following sections detail how these steps were performed.

Declaration of Grain

The chronological processes of installing ARUs, the acquisition of acoustic recordings and the identification of species from these recordings (together with their respective granularity of deployment, recording and detection) were adopted as research activities for which individual data marts were made (Table 3.9).

Table 3.9: Major activities of bioacoustic research, granularity of data and the designated data mart.

Research Activity	Description	Granularity	Data Mart
Installation of ARUs	details of where and when ARUs were placed	Deployment	Deployment Mart
Creation of Acoustic Recordings	details pertaining to the acquisition of bioacoustic recordings from deployed ARUs	Recording	Recording Mart
Identification of Species	determination of animal type based on a recorded acoustic event	Detection	Detection Mart

Identify the Dimensions

A procedure called the “business bus matrix” (Kimball et al., 2002) was used to identify which dimensional tables (DDP Step 3) were required by each fact table of each data mart. Knowing which data marts will share a dimension table ensured that these tables were designed to conform across the relevant data marts (Table 3.10). Because bioacoustic research (not business) activities were used, the process will hereafter be called only the “data warehouse bus matrix”.

Table 3.10: Data warehouse bus matrix. Granularity of research activities employed in bioacoustic research.

		Dimensions				
Activity	**Granularity**	**Date**	**Study**	**Habitat**	**Acoustic File**	**Species**
Install ARUs	Deployment	✓	✓	✓		
Collect recordings	Recording	✓	✓		✓	
Identify species	Detection	✓			✓	✓

The purpose of each dimension table is defined as follows:

Date: the year, month and day on which an event or activity occurred.

Study: the association ARUs have to each other in the context of research projects

Habitat: measured and assessed characteristics of an ARU deployment site.

Acoustic File: information about the digital file recorded by an ARU.

Species: the taxonomic name of an organism identified from an acoustic event recorded by an ARU.

Identify Facts

Quantifiable values, as they were related to each research activity, were copied to their respective fact tables.

Time, was considered at two scales: days and time–of–day. At the scale of days, time was considered a nominal attribute and was represented by the Date dimension. At the scale of time–of–day, it was considered as interval data and stored as a fact.

Results

The results of the data processing through the steps of the Knowledge Discovery (KDD steps 2 and 3) in Databases and the creating of a data warehouse are presented in this section.

Knowledge Discovery in Databases Step 2 – Processing of Data to Handle Missing and Erroneous Data

The method by which missing and erroneous data were resolved is outlined for the attributes: dates, locations, species, water depth, wetland assessment, horizontal cover.

Dates

Of the total 7,933 ARU deployment records downloaded, missing or erroneous dates were found in a total of 685 (19%) in the following categories:

Class 1 Missing Values (null) By selecting only records deployed before 2015, 17/7,933 (0.2%) records with null deployment dates were excluded. Retrieval dates were found to be null in 665/3,533 (19%) of records.

Class 2 Out-of-Range One record out of 3,533 (0.03%) was found to have the unrealistic deployment date of 1960/08/01.

Class 3 Out-of-sequence 9/3,533 (5%) records had retrieval dates that precede deployment dates as shown in Table 3.11.

Class 4 Range Greater than one year A review of the length of time ARUs were deployed revealed that 179/3,533 (5%) records had deployment dates greater than 365 days (Table 3.12). Since it is plausible that ARUs were deployed for one full year (and that some in remote locations could were left until retrieval was convenient) a threshold of greater than 1.5 years was chosen to identify unrealistic deployment lengths, leaving 10/3,533 (0.03%) records.

Records which were in one of these four error classes were identified as having invalid or missing date values by setting their IS_COMPLETE attribute to 0.

Table 3.11: Records where retrieval dates precede deployment dates

Deployment Date	Retrieval Date	Deployment Length (days)
2014-05-14	2004-05-27	-3,639
2014-06-23	2014-06-09	-14
2012-05-24	2012-05-19	-5
2014-06-24	2014-06-20	-4
2014-06-14	2014-06-12	-2
2014-05-30	2014-05-29	-1
2014-06-13	2014-06-12	-1
2014-06-16	2014-06-15	-1
2014-06-16	2014-06-15	-1

Table 3.12: Occurrence of ARU deployments exceeding 1 year in length. Records marked in bold are considered to be erroneous.

Deployment Length (days)	Number of deployments
19,686	**1**
6,035	**1**
6,010	**2**
4,974	**2**
4,973	**2**
4,397	**1**
4,027	**1**
411	9
410	5
367	1
366	100

Location

Deployment records were eliminated if latitude and/or longitude was not recorded and for those points not within the Lower Athabasca Region. A total of 228/3,533 (6%) had

incomplete coordinate data. The remaining 3,305 were imported into the GIS where those within the study region were extracted by an overlay with the Lower Athabasca Region polygon and imported back into the Oracle database. Those 2,077 deployment records were then marked with a 1 in the IN_REGION attribute.

Species

Of the 122,743 records of species detections, 478 (0.5%) did not match the species codes used by the U of A researchers. One species code (YWAR) was found which did not match the list of possible codes. The code was recognized by this researcher to be the old code from Yellow Warbler (*Setophaga petechia*) and has since been changed to YEWA. This species code was corrected for the 443 (0.4%) records where it occurred. Three more erroneous entries, corresponding to 10 (0.01%) records, could be reasonably attributed to the correct codes. The remaining 25 (0.02%) were excluded from the database (Table 3.13).

Table 3.13: Records of species determinations from ARU recordings where the code entered by the listener does not match the list used by the University of Alberta. Codes which could be corrected are marked as bold.

Unmatched Species Code	Correct Species Code	Description	Occurrence
NSSP	—	—	7
NTRLL	—	—	1
TTWO	—	—	6
UNBUTEO	**UNBU**	**Unknown Buteo**	**1**
UNCH	—	—	1
UNCR	—	—	10
UNGO	—	—	2
UNKI	—	—	1
UNTERN	**UNTE**	**Unknown Tern**	**8**
UNTRL	**UNTRLL**	**Unknown Trill**	**1**
UNTTBB	—	—	3
YBWO	—	—	1
YWAR	**YEWA**	**Yellow Warbler**	**443**
Total:			478

Within the species list used in the EMCLA database, the codes LABU, LALO and SMLO were each used redundantly (Table 3.14). The code LABU was mistakenly used as the code for Lazuli Bunting (*Passerina amoena*) and Lark Bunting (*Calamospiza melanocorys*) which should have LAZB and LARB respectively. Records for the codes LALO and SMLO were correctly used for Lapland Longspur and Smith's Longspur, each in the family *Calcariidae* but duplicate records for these species occurred under the old family, *Emberizidae*.

Table 3.14: Redundant codes (LALO, SMLO and LABU) found in the EMCLA species list.

Family	Genus	Species	English Name	Redundant Species Code
Cardinalidae	*Passerina*	*amoena*	Lazuli Bunting	LABU
Emberizidae	*Calamospiza*	*melanocorys*	Lark Bunting	LABU
Emberizidae	*Calcarius*	*lapponicus*	Lapland Longspur	LALO
Calcariidae	*Calcarius*	*lapponicus*	Lapland Longspur	LALO
Calcariidae	*Calcarius*	*pictus*	Smith's Longspur	SMLO
Emberizidae	*Calcarius*	*pictus*	Smith's Longspur	SMLO

The current species codes and family designations were included in the data warehouse. No detection records were found which contained the ambiguous code LABU, so no corrections were required. The records containing the *Emberizidae* family of Smith's Longspur and Lapland Longspur were excluded from the DSS.

Water Depth

Water depth data was available for 736/7,933 (9%) of ARU deployments. Of the 18,787 measurements taken, 165 (0.9%) contained water depth values outside of the acceptable range (Table 3.15) and 5 (0.03%) contained a null value. These values are most likely "sentinel values", used to convey a message and were excluded from the study. Subsequent calculations were made using the remaining, valid values. The 18,616 records were used in subsequent aggregation calculations.

Table 3.15: Water Depth measurements out of the expected range or missing.

Depth Recorded	Occurrence
99999	12
9999	15
999	18
-5	3
-881	9
-882	88
-883	13
-884	7
null	5
Total	170

Wetland Assessment

Habitat assessments were available for 499/7,933 (6%) of the ARU deployment sites, comprising 2,052 individual records. Problems were encountered where field technicians deviated from the standard DUEWC codes. Four different error classes were noted: transposition of the code occurred in 6/2,052 (0.3%) records, only partial information about the habitat was recorded in 48/2,052 (2%), more than one habitat class was recorded for the same area in 14/2,052 (0.7%) and completely invalid information was recorded in 120/2,052 (6%) (see Appendix Table A1)s. Additionally, 19/2,052 (0.9%) contained null values for habitat.

Transposed entries were simply classified to their intended categories but accommodating partial classifications and novel classifications required adjustments to the way the DUEWC system was represented in the database. Entries that specified more than one class were omitted because no valid decision could be made on the importance of either.

Because the DUEWC system is a hierarchical classification, a category called "Undefined" was included at each level to accommodate entries which did not specify all levels. Additionally, a category called "Undefined Fen" was included for records which did not specify a fen's nutrient type and the Minor Wetland Classes named

“Anthropogenic Disturbance” and “Natural Disturbance” were added to record the noted occurrence of habitat alterations such as fires and cutlines.

Other modifications were the elimination of the the Major Soil Group, as it was never directly appraised, and the promotion to the Major Wetland Class of the categories “Undefined Fen”, “Rich Fen” and “Poor Fen”, which exist as their own level between the Major and Minor Wetland Classes. Figure 3.6 shows the modifications made to the DUEWC classes.

A translation table was used to associate each non-standard DUEWC classes to the appropriate hierarchical level in the database. Table A2 in the Appendix shows how the translation table matched the non–standard habitat entries with auxiliary codes and employed the added classes “Undefined”, “Undefined Fen”, “Anthropogenic Disturbance” and Natural Disturbance.“ A total of 62/2,052 (3%) of the non-standard habitat codes could not be included in the database.

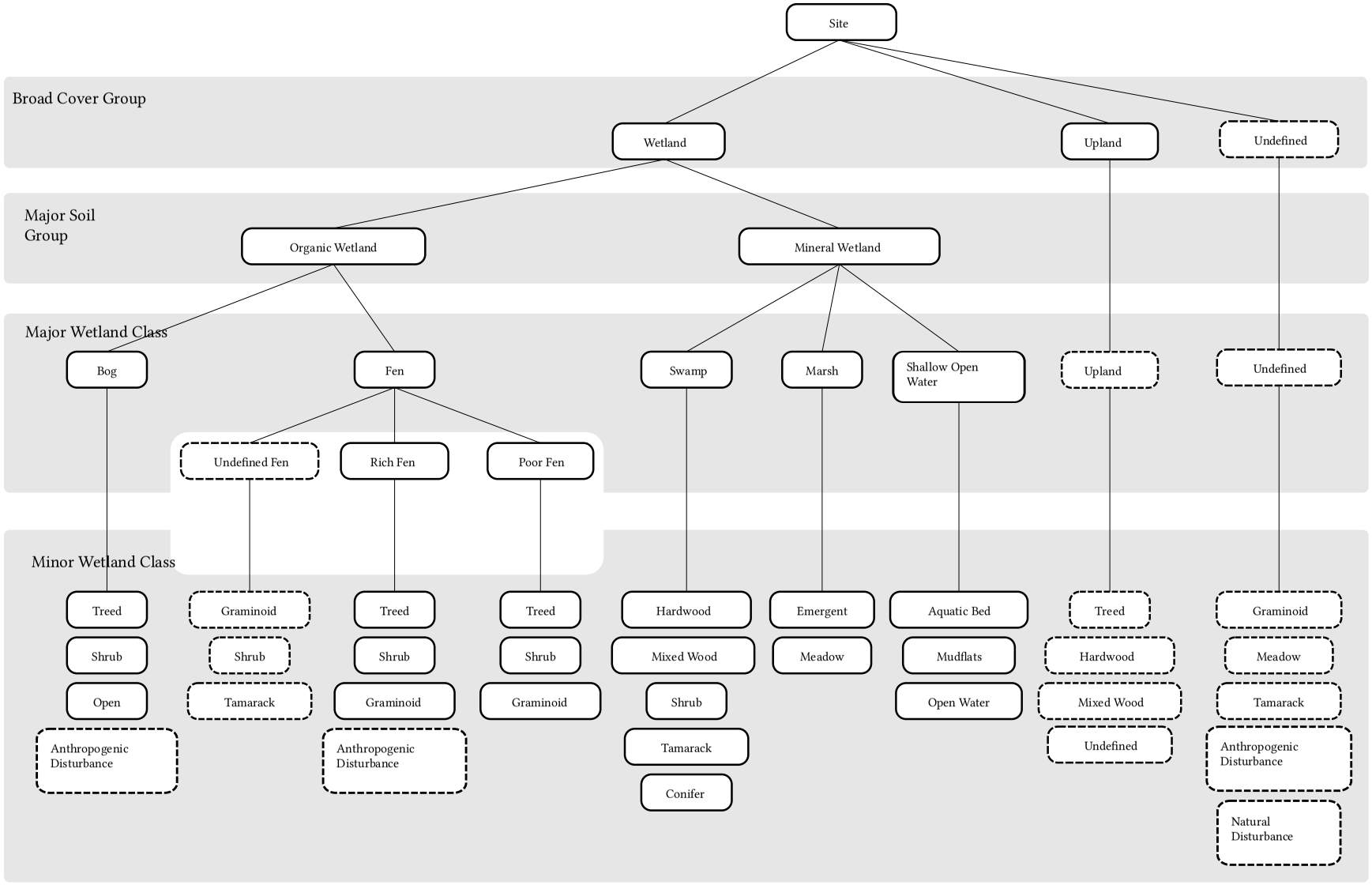


Figure 3.6: Modified Ducks Unlimited Enhanced Wetland Classification System showing the addition of new categories (dashed boxes) and the promotion of Rich Fen and Poor Fen to the Major Wetland Class.

Horizontal Cover

Horizontal cover was available for 499/7,933 (6%) of ARU deployment sites, comprising 5,067 individual records . Of these, 57/5,067 (1%) records had values outside of the expected range of 0% to 100% (Table 3.16). These were most likely sentinel values and were excluded from calculations of average cover.

Table 3.16: Horizontal cover estimates.

Cover Recorded	Occurrence
9999	28
666	10
-5	5
-82	14
Total:	57

Knowledge Discovery in Databases Step 3 – Reducing Dimensionality

Acoustic Recordings

Due to insufficient computer memory to analyze a full ten-minute stereo recording, each file was subdivided into 1-minute-long sections and only the left stereo channel was analyzed. At a bitrate of 44,100 kHz, each one minute recording had 2,646,000 $\mathrm{samples}$ so a PAA number of 264,600 was used for a tenfold reduction (Table 3.17).

The size of the left stereo channel of the 10 minute recording was $52.9\,\mathrm{MB}$, while the file created by the PAA/SAX reduction was $5.3\,\mathrm{MB}$ (Table 3.17).

Table 3.17: Parameters for PAA/SAX reduction of a ten-minute long bioacoustic wav file recorded by ARU (left channel only).

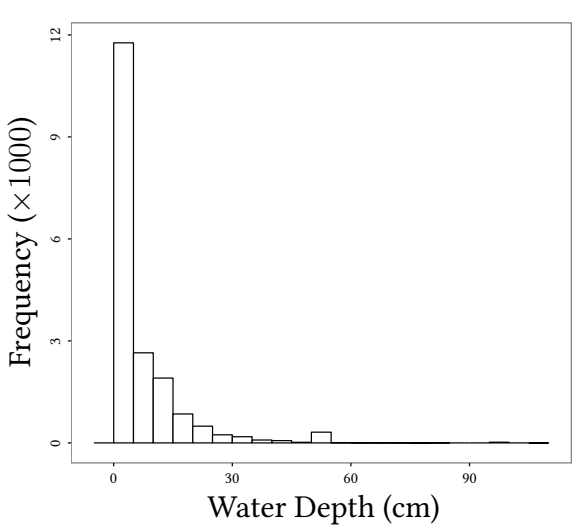
Parameter	**Size**
alphabet size	8 characters
bitrate	44,100/sec
samples in 1 minute	2,646,000
PAA (at 10 samples long)	264,600
File size (.wav format)	52.9 MB
PAA/SAX file size	5.3 MB

Water Depth

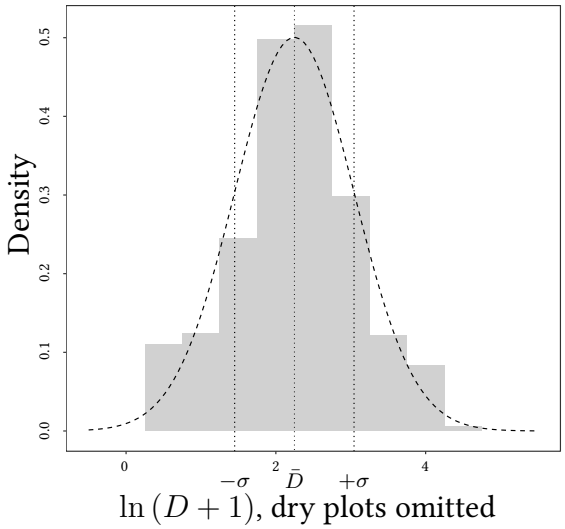
The highly skewed water depth values for wet plots (Figure 3.7a) were transformed by taking the natural logarithm of each value plus 1 (Equation 3.2) to create an approximately normal distribution (Figure 3.7b). The approximately straight line formed by a Q-Q plot confirmed a nearly normal distribution (Figure 3.8).

The value of one was added to ensure that all transformed values were positive.

$$d' = \ln(d + 1) \tag{3.2}$$



(a) Frequency of water depth values (cm) recorded at recorder deployment sites for the Yellow Rail study.



(b) Transformed water depth (excluding dry plots), vertical lines show average water depth $(\bar{D})$ and 1 standard deviation (σ) of the distribution.

Figure 3.7: Frequency of raw and transformed depth values recorded at some ARU deployment sites.

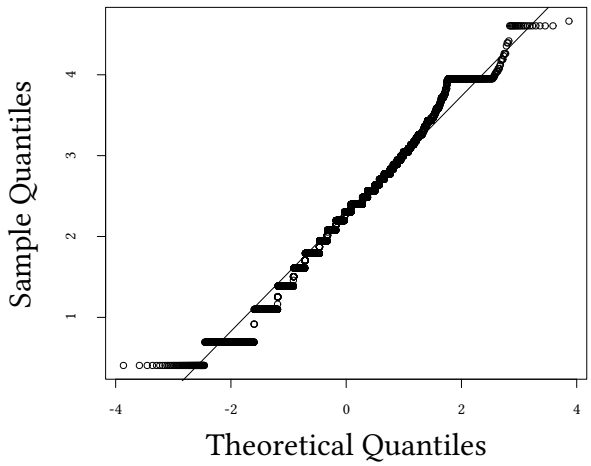


Figure 3.8: QQ plot of transformed water depth showing a nearly normal distribution.

Three fuzzy membership sets for water depth were derived from the transformed distribution following Suh (2012). The mean, minimum, maximum and standard deviation ($\bar{D}$, D_{min}, D_{max} and σ) were calculated and used to derive the values of

$\bar{D}_{shallow}$ and $\bar{D}_{deep}$ for the deep and shallow subsets of the of transformed data using Equation 3.3 and Equation 3.4 (Table 3.18).

$$\bar{D}_{shallow} = \frac{1}{n}\sum_{i=1}^{n}\left\{d : d < \bar{d} - \sigma\right\} \tag{3.3}$$

$$\bar{D}_{deep} = \frac{1}{n}\sum_{i=1}^{n}\left\{d : d > \bar{d} + \sigma\right\} \tag{3.4}$$

Table 3.18: Water depth values used to calculate fuzzy depth membership functions.

Value	Transformed Depth (ln(*d*+1))	Untrainsformed Depth (cm)
$\bar{D}_{min}$	0	0
$\bar{D}_{shallow}$	1.06	1.89
$\bar{D} - \sigma$	1.45	8.46
$\bar{D}$	2.25	3.26
$\bar{D} + \sigma$	3.04	19.99
$\bar{D}_{deep}$	3.48	31.44
$\bar{D}_{max}$	4.66	105

These values were then "un-transformed" by raising the mathematical constant e to the power of each transformed value (d') and adding one (Equation 3.5).

$$d = 1 + e^{d'} \tag{3.5}$$

The un-transformed values were used to define three triangular fuzzy sets (shallow, medium and deep) (Figure 3.9) which were used to calculate the fuzzy memberships (μ) for all un-transformed depth values at each ARU deployment (including dry plots, with depth values of $0\,\mathrm{cm}$). An arithmetic mean was also calculated.

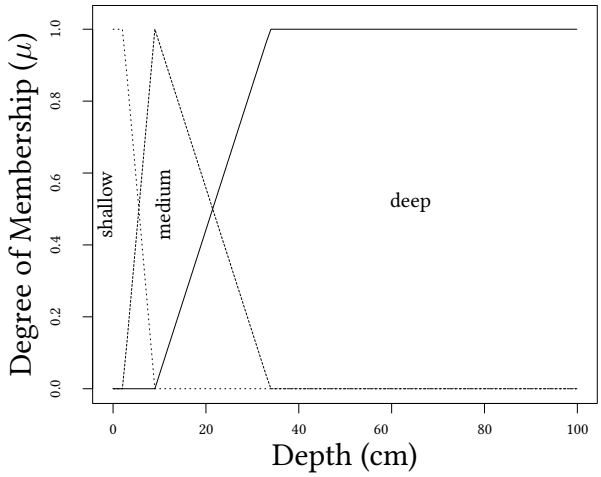


Figure 3.9: Membership functions for un-transformed water depths

Thus, for each station water depth was reduced from 21 individual measurements to four aggregate values: three fuzzy memberships and one simple mean. Figure 3.10 shows how an instance where fuzzy memberships reflect the distribution of depth measurements better than the arithmetic mean. The shaded portions of the fuzzy sets (A) indicate higher Shallow and Deep memberships than the Medium–depth membership which reflects the distribution of depths recorded (dots in figure B). In contrast very few depth measurements are similar to the arithmetic mean, shown as a dashed line in (B).

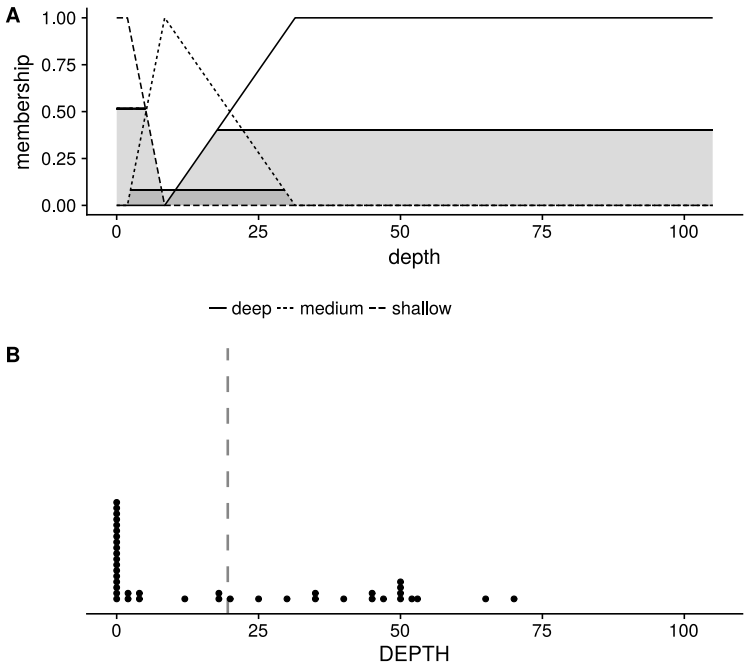


Figure 3.10: Water depth aggregation for a single ARU Station calculated using fuzzy logic and by arithmetic mean. Three fuzzy sets (A) for Shallow, Medium and Deep are shown with membership values for each (shaded areas) Shallow = 0.51, Medium = 0.09 and Deep = 0.40. The calculated arithmetic mean (B) of $19\,\mathrm{cm}$ is shown by a dashed line. The distribution of depth memberships is indicated by dots in (B). The horizontal axes of both graphs are scaled the same to allow comparison.

Horizontal Cover

The ten horizontal cover measurements collected at each ARU deployment were aggregated to three arithmetic means, for low, high and combined cover.

Wetland Classification

The area of each class was calculated as a proportion of the total buffer area, based on the area of the inner buffer (A_1) and the ring-shaped out outer buffer (A_2) which were defined as:

$$A_1 = \pi r_1^2 \tag{3.6}$$

$$A_2 = 8\pi r_1^2 \tag{3.7}$$

where $r_1 = 50\,\mathrm{m}$ and given that $r_2 = 150\,\mathrm{m} = 3r_1$.

Although the two habitat value were reduce to a single value, the inner and outer buffer data were still included in the database since they reflect the proximity of the habitats surrounding the ARU deployment.

The cleaned and reduced data was then taken through the steps of the Dimensional Design Process to create a data warehouse.

Implementing the Dimensional Design Process

Identify the dimensions

The dimension tables identified previously through the data warehouse bus matrix (Table 3.10) were Date, Study, Habitat, Acoustic File and Species. How each was implemented is described below.

Date Dimension The date dimension, called DIM_DATE, was created to cover the range of dates from 2000/01/01 to 2050/01/01 and included several fields to enhance its utility. The table was created to a distant future date but could be easily be extended beyond that if the DDS is still in use by then. Table 3.19 shows the names, the formats, ranges and suggested uses of the fields in the date dimension.

In addition to recording the standard (Gregorian) date recorded for an event, fields were included to parse dates at the resolution of year, month and day, as well as count of Julian Days Number (which is designated to have started at January 1, 4713 BCE). Additionally, to assist in time calculations, a flag indicating daylight savings time and the UTC offset for Mountain Time (corrected for Daylight Savings Time), were included.

The table can be extended beyond the current final date if required and future changes to the application of daylight savings time can be incorporated through manual adjustments.

Table 3.19: Features of the date dimension (DIM_DATE).

Column Name	Format	Range	Suggested Use
EVENT_DATE	YYYY/MM/DD	2000/01/01 to 2050/01/01	To determine the chronological order of events.
EVENT_YEAR	YYYY	[2000, 2050]	To group events by field-season.
EVENT_MONTH	MM	[1, 12]	To identify frequency patterns of events at a monthly resolution, within a year.
EVENT_DAY	DD	[1, 31]	To identify frequency patterns of events at a daily resolution, within a month.
SINCE2000	Whole Number	[1, 18 264]	To calculate the number of days between the occurrence of different events, within this database.
DAYOFYEAR	Whole Number	[1, 365] normal years [1, 366] on leap years	To identify frequency patterns of events at a daily resolution, within a year.
JULIAN_DAY	Whole Number	[2 451 545, 2 469 808]	To calculate the number of days between the occurrence of different events, between any databases using this standard.
IS_DST	A 1/0 flag indicating if the day is a daylight savings date	[0, 1]	To identify events occurring at standard time or daylight savings time.
UTC_OFFSET	Mountain Time zone offset from Coordinated Universal Time, adjusted for daylight savings time.	[6, 7]	To make comparisons between events in different time zones or between standard and daylight savings time

Study Dimensions A dimension named DIM_STUDTY was prepared to contain the hierarchical categories of Project Name, Site and Station by which ARU deployments are logically grouped (Table 3.20).

Table 3.20: Dimension for ARU deployment data.

Field Name	Description
ID_DIM_STUDY	A unique identifier for records in this table
PROJECT_NAME	A collection of activities performed to answer a research question
STATION	A grouping of related research activities
SITE	A distinct location where research activities occurred

Habitat Dimension Because each ARU deployment could be associated with many DUEWC habitat classes, and each habitat class could be associated with multiple ARU deployments, a many–to–many relationship existed. To accommodate this condition, a linking table called DEPLOYMENT_WETLAND_MEMBERSHIP was created. Each row of this table contained the identification number (ID) of a deployment record, the ID number of a DUEWC habitat class and the proportion of the habitat within the inner buffer, outer buffer and both buffers combined. Table 3.21 shows how an ARU deployment with a ID_DEPLOYMENT_FACT of 11 is associated with both ID_WETLAND_CLASSs 1 and 2, while ID_WETLAND_CLASS 1 is associated with ARU deployments with ID_DEPLOYMENT_FACT value of 111 and 129.

Table 3.21: Example data contained in the DEPLOYMENT_WETLAND_MEMBERSHIP, showing how the many–to–many link between an ARU deployment and a DUEWC wetland class is represented by allowing multiple occurrences of values in the ID_DEPLOYMENT_FACT and ID_WETLAND_CLASS fields.

ID_DEPLOYMENT_FACT	ID_WETLAND_CLASS	INNER_BUFFER	OUTER_BUFFER	WHOLE_BUFFER
111	1	1.0	0.6	0.64
111	20	(null)	0.1	0.09
129	1	0.95	0.95	0.95

Dimensional reduction of the wetland classification data was accomplished by collapsing the hierarchical structure of the DUEWC system into a single “flat file” which was used to construct the dimension table DIM_DUE_WETLAND_CLASSIFICATION.

Table 3.22 shows a schematic diagram of the DUEWC dimension table. Hierarchical levels of the system are represented with two modifications. The Major Soil Group was excluded because it was not directly assessed by field technicians and is not applicable to Upland habitats. The Minor Wetland Classes of Rich Fen and Poor Fen were promoted to the Major Wetland Class level, replacing the Fen category (Figure 3.6).

Table 3.22: Schematic representation of the hierarchical structure of the Dimension for Ducks Unlimited Enhanced Wetland Classification (DUEWC) assessments used to populate the DIM_DUE_WETLAND_CLASSIFICATION dimension table. Added categories are shown in italics.

Cover Group	Major Wetland Group	Minor Wetland Class
Undefined	*Undefined*	*Anthropogenic Disturbance*
		Graminoid
		Meadow
		Natural Disturbance
		Tamarack
Upland	*Upland*	*Deciduous*
		Mixedwood
		Treed
		Undefined
Wetland	Bog	*Anthropogenic Disturbance*
		Open
		Shrubby
		Treed
	Marsh	Emergent
		Meadow
	Poor Fen	Graminoid
		Shrubby
		Treed
	Rich Fen	*Anthropogenic Disturbance*
		Graminoid
		Shrubby
		Treed
	Swamp	Conifer
		Hardwood
		Mixedwood
		Shrubby
		Tamarack
	Undefined Fen	*Graminoid*
		Shrubby
		Tamarack

Acoustic File Dimension A table called DIM_FILENAME was created for the acoustic dimension, containing a unique identifier field and a text field for the file name.

Species Dimension A table called DIM_SPECIES was created for the species dimension (Table 3.23). This table contained the attributes for the taxonomic levels from order to species, common names in English, the standard 4-letter code used by the American Ornithological Union (North American Classification Committee, 2014) and logical fields for the species occurrence in Alberta and its vocalizations characteristics.

Table 3.23: Attributes of the species dimension table.

Field Name	Description
ID_DIM_SPECIES	A unique identifier for each record
TAX_ORDER	Taxonomic order
FAMILY	Taxonomic family
GENUS	Taxonomic genus
SPECIES	Taxonomic species
ENGLISH_NAME	Non-technical English name for a animal
CODE	Standard American Ornithological Union 4-letter codes plus EMCLA codes
IN_ALBERTA	Boolean (Yes/no) indicating the natural occurrence in Alberta
SONG_VOCAL	Boolean (yes/no) indicator if the animal vocalizes a song
CALL_VOCAL	Boolean (yes/no) indicator if the animal vocalizes a call
NO_VOCAL	Boolean (yes/no) indicator if an animal creates a non-vocal sound

Identify the facts

Fact tables for each of the research activities identified in Table 3.10 were created and loaded with the the corresponding data. The contents of the fact and dimension tables for each data mart are described below, followed by a diagram of each.

Deployment Mart This data mart was created to contain the temporal, spatial and habitat context associated with the installation of an ARU.

Numeric attributes included in the DEPLOYMENT_FACT table were: geographic coordinates, water accumulation, mean horizontal cover, mean water depth, memberships in the fuzzy water depth sets and the proportional of each CanVec+ habitat category around the ARU deployment site.

The following relationships were made to dimension tables: DIM_STUDY, DIM_DATE (one each for deployment and retrieval of the ARU) and DIM_DUE_WETLAND_CLASSIFICATION through the DEPLOYMENT_WETLAND_MEMBERSHIP linking table (Figure 3.11).

Recording Mart This data mart was created to contain temporal and metadata associated with the acquisition of bioacoustic recordings from ARUs.

The attributes contained in the RECORDING_FACT table were: the time when the recording started, the number of minutes before and after sunrise, the number of minutes before and after sunset and the symbolic representation of the recording.

Relationships were made to the dimension tables: DIM_DEPLOYMENT, DIM_FILENAME and DIM_DATE (Figure 3.12).

Detection Mart This data mart was created to contain the temporal and taxonomic information associated with the identification of species from bioacoustic recordings.

The DETECTION_FACT table contains only links to the dimension tables: DIM_DATE, DIM_FILENAME and DIM_SPECIES (Figure 3.13).

All data marts combined to form the full data warehouse as show in the Appendix, Figure A1.

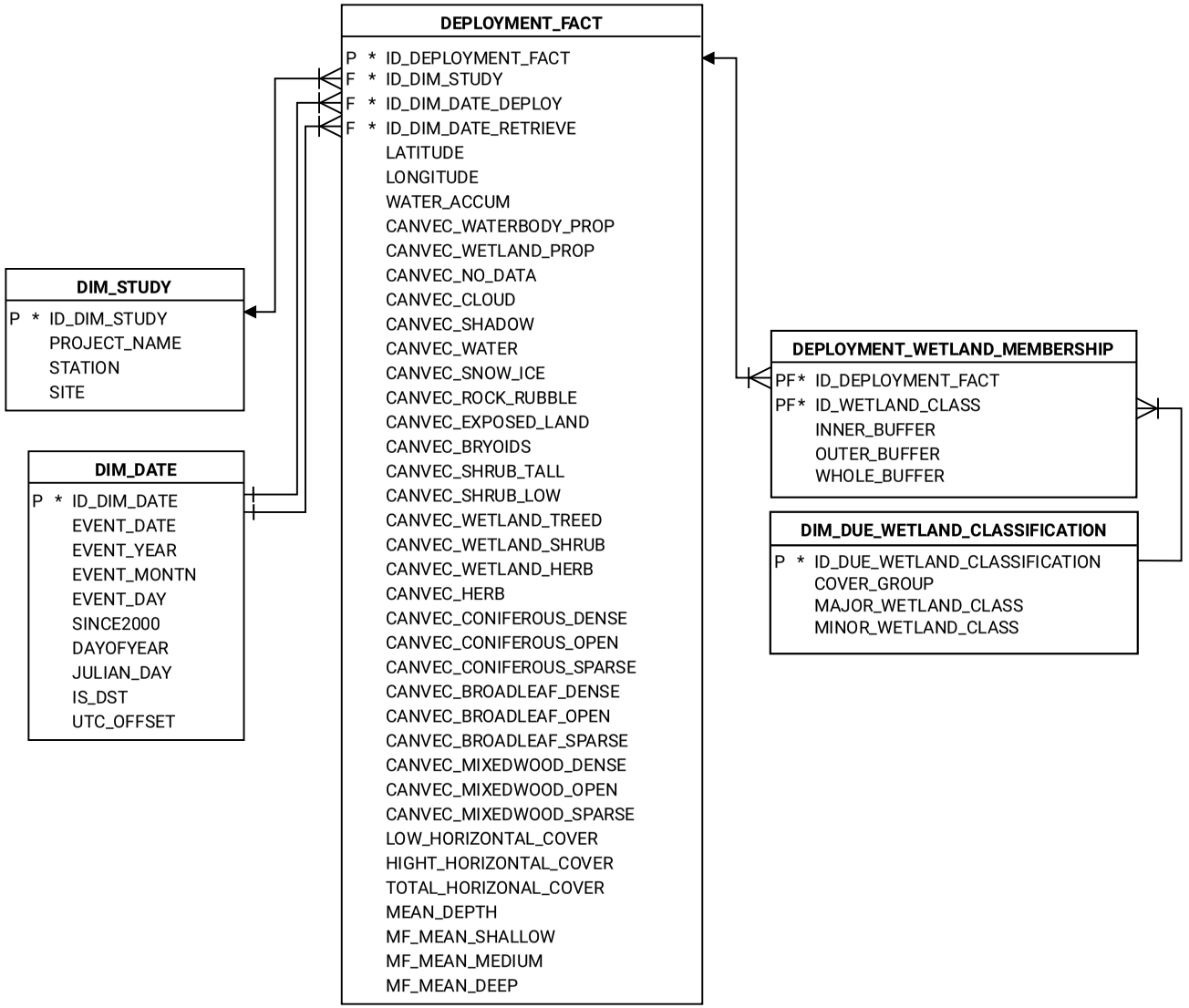


Figure 3.11: Deployment Mart: a data mart containing information related to the deployment of ARUs.

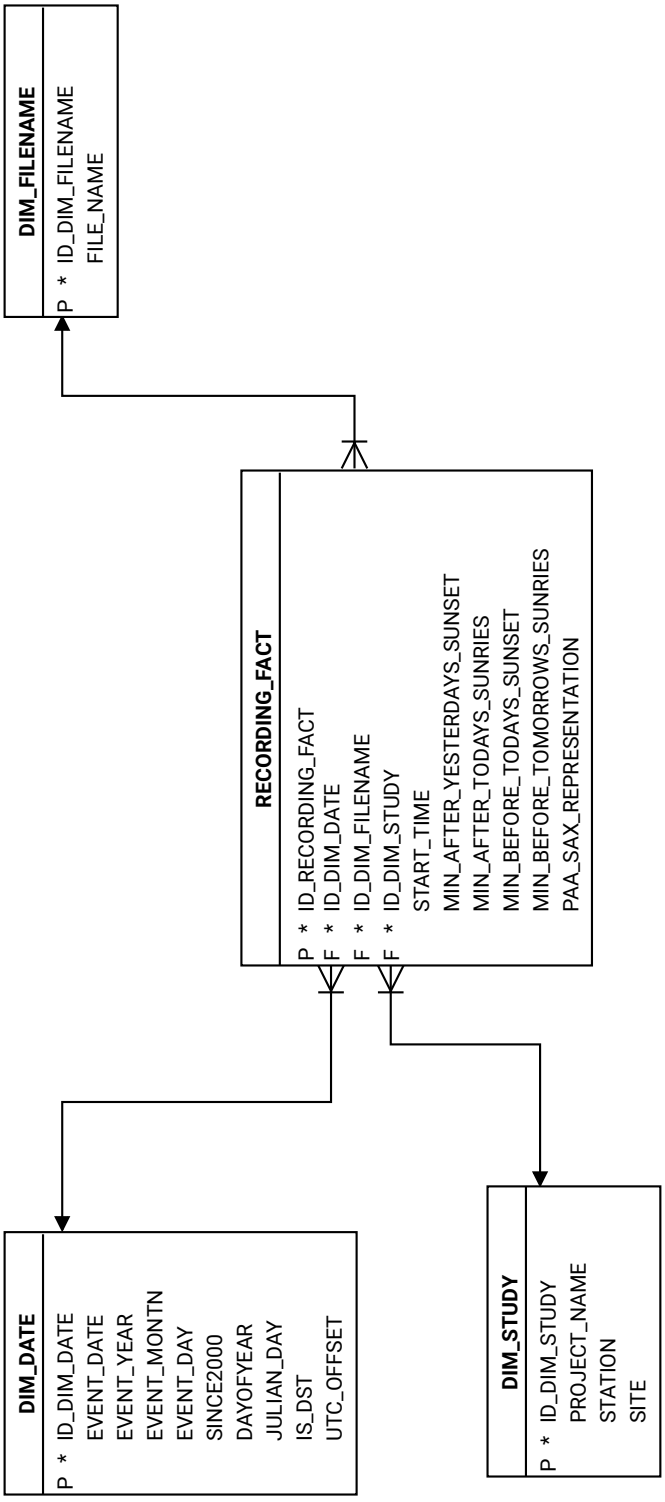


Figure 3.12: Recording Mart: a data mart containing information related to recordings made by ARUs.

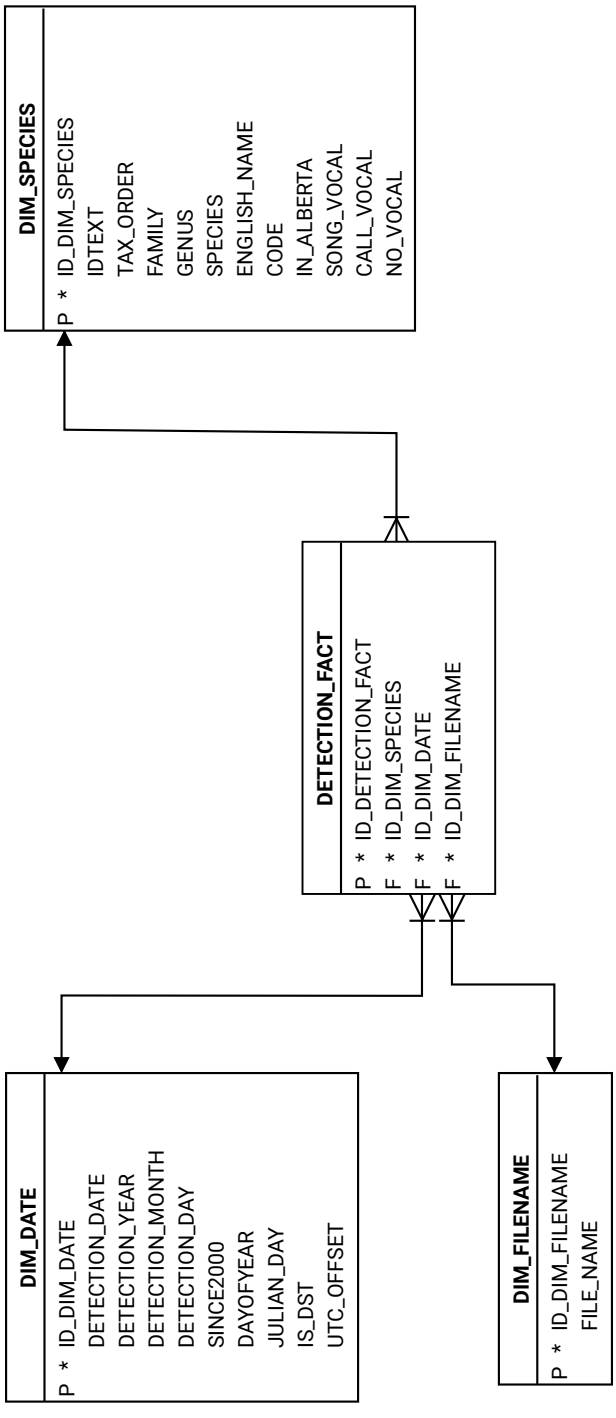


Figure 3.13: Detection Mart: a data mart containing information related to the species detected from ARU recordings.

Discussion

This research demonstrated the creation of a data warehouse for an existing set of bioacoustic research data following the Knowledge Discovery in Databases (KDD) procedure and the Dimensional Design Process (DDP). The data warehouse was composed of three data marts, created to house data associated with three research activities: deployment of ARUs (Deployment Mart), creation of bioacoustic recordings (Recording Mart) and the identification of species from these recordings (Species Mart). These data marts were related to each other through shared dimension tables (Figure A1). The data warehouse was created as the starting point of a Decision Support System for bioacoustic data processing. The data warehouse could equally be used to facilitate exploration of the data to find patterns which could inspire new research. The remainder of this section discusses the issues and advantages encountered in creating the data warehouse.

A key issue in creating the data warehouse was problematic records encountered in the data. These result from the inherent difficulties in collecting ecological data under field conditions, measuring ecological components and from the inexperience of field technicians. The presence of these inconsistencies highlight the need to seek and address missing and erroneous data (KDD step 2). A thorough understanding of the ecological components being characterized and the techniques used to measure them must be possessed to assess the validity of each attribute. In this research data was validated by a comparison to the domain of values expected for the instrument or techniques used (e.g.: $0\,\mathrm{cm}$ to $100\,\mathrm{cm}$ for water depth measured with a meter stick) but also by comparison to the values of related attributes (e.g.: confirmation that each ARU's deployment date precede its retrieval date).

Attempts were made to glean as much information as possible from non–standard data because field data is expensive to collect and often represents an ephemeral set of conditions. Knowledge of the data was necessary to make decisions on how erroneous data should be handled. Where errors impacted the certainty of the time and location of a species detected from a recording, those deployment records were omitted. But, in circumstances where enough data remained to serve the original purpose, only problematic attributes were omitted. This was the case for deployment sites where some water depth measurements were invalid, the remaining valid measures were used to characterize that ecosystem attribute. Without an understanding of the data, judgments such as these would not be possible.

Errors were also found where judgment was required to categorized ecosystems within an established classification scheme, in this case the Ducks Unlimited Enhanced Wetland Classification (DUEWC) system. These ranged from simple transpositions of the conventional codes, to entries indicating only part of a habitat class, to entries with indicated multiple classes simultaneously, to entries indicating conditions not classified in the DUEWC system. In cases where ambiguity was present, the records were omitted. Inclusion of many non–standard entries in the data warehouse was made possible by the way the DUEWC system was represented within the data. By treating each class of the hierarchical system as an attribute, the known part of incomplete entries could be assigned a value while the missing class could be assigned a new value called "Unknown". For example, rather than recording a site as having a DUEWC class Tamarack Swamp (STM), all levels of the hierarchy are specified as follows: Broad Cover Group = "Wetland", Major Wetland Class = "Swamp" and Minor Wetland Class = "Tamarack". By the addition of class values not encompassed in the DUEWC system (e.g.: "Anthropogenic Disturbance") more non–standard data were included.

In some instances attribute values were found to be so far outside the expected domain they must represent a attempt by the field technician to communicate why a certain measurement could not be taken using a "sentinel values". The presence of sentinel values indicates a need to allow field staff to record when a value could not be measured (which should be entered as a NULL value in the data base). When designing protocols for the collection of field data, researchers should anticipate as far as possible, conditions where data collection may not be possible or the ecological attribute may not conform to expectations. Procedures and data collecting interfaces (computer or paper based) should be designed to allow field staff to record this information in ways which do not contaminate data with "sentinel values". Interfaces should also be designed with validation checks to prevent collection of data which does not comply to the domain of possible values.

One advantage of transforming date to include in a data warehouse is the opportunity to choose techniques which can best represent the meaning of the data through aggregation (KDD step 3). In this research, water depth measurements were taken at 21 points around some ARU station in order to characterize the surrounding wetland. While a simple arithmetic mean was calculated, fuzzy logic was also used to derived a more nuanced description of the habitat. There is a great deal of plasticity in the technique, for instance there is no "correct" way to create the fuzzy sets used. In this research the distribution of data was used to derived the sets meaning that membership of a site in

each set is relative to all sites. The fuzzy sets could also have been defined to reflect expert knowledge, such as the preferred habitat of a species of interest. The decision of how to apply fuzzy logic to the data is where ecologists are able to express the ecological "meaning" of the data in a way which best represents the ecosystem component.

Similar to aggregation, managing digital files in a way which facilitates machine processing is a challenge increasingly faced by ecologists as more affordable data loggers become available. Time–series data are produced by many other devices now being employed in ecology, such as light–level archiving "geolocators" and miniature "iButton" environmental monitors (Dallas Semiconductor – a subsidiary of Maxim Integrated Products, Sunnyvale, California). Making these data searchable in a data warehouse greatly increases an ecologists ability to analyze their results. In this thesis PAA/SAX was chosen to reduce bioacoustic ARU files to a series of characters. Different techniques to reduce time–series signals may be suited to different data or for different interpretations. However, an ecologist does not have to chose one reduction technique but can store multiple approximation of their time series.

Another advantage of using a data warehouse is the ability to easily include data from other sources and by calculation. For example, habitat classifications and water accumulation measures were extracted from a GIS while recording times were calculated relative to sunrise and sunset. In the first case, data which was impractical to measure from ground surveys was achieved from thematic maps. In the second case time, as perceived by the organisms under study, was made available for analysis. The data mart structure of the Kimball (1996) data warehouse allows additions of derived data through adding attribute to the Fact or Dimension tables, by adding new Dimension tables or by adding new data marts.

How the data warehouse created in this thesis could be applied to the proposed DSS, as well as considerations for the general use of data warehouses in ecology, are presented in the next chapter.

Chapter 4

Conclusion

This research was inspired by the author's observations on the changing field of avian monitoring, where Autonomous Recording Units (ARUs) have been replacing surveys conducted by expert human "birders". While expanding the number of avian surveys conducted, ARUs have introduced the burden of processing large numbers of bioacoustic files. Expert human "listeners" are still required to identify species from the recorded vocalizations, although there is an active effort aimed at fully automating this process. To help listeners keep pace with the production of bioacoustic recordings, a Decision Support System (DSS) was suggested which would identify likely species based on acoustic features and recording context. A major component of the DSS, a data warehouse, was created in this research.

The contribution of this thesis was to outline the creation of a data warehouse for a set of bioacoustic data in order to illustrate how a data warehouse can be built from existing ecological data which was not collected for this purpose. Although this data warehouses was designed as a component of a DSS, the utility of this data structure in accessing large amounts of data from multiple sources makes it useful by itself. Data warehouses are not commonly used by ecologists, though it can be argued that they are more suited to the research field than to the business world for which they were developed. In a business case, the primary activity is to conduct financial transactions, while the analysis of archived data is an important but secondary task. Conversely, ecologists and other researchers collect data primarily to be analyzed.

Two other techniques not widely used in ecology were also described. First, fuzzy logic, which is used mostly in engineering, was used to represent the imprecise habitat attribute of "wetness" at ARU sites. Second, Piecewise Aggregate Approximation and

Symbolic Aggregate approXimation (PAA/SAX), which were developed in the field of computing science to reduce time series data, were used to improve the searchability of bioacoustic recordings. These techniques are useful to represent, in a data warehouse, two common types of ecological data: imprecise data due to the difficulty of measuring and categorizing ecological components and time series data generated by data loggers.

The utility of the data warehouse as part of the proposed DSS can not be tested until the rest of the DSS is created and its effectiveness at assisting human listeners to process bioacoustic recordings can be assessed. This would require not only an evaluation of how easily the desired attributes could be retrieved but also whether the attributes stored would be sufficient to allow the DSS to make useful species suggestions. However, the value of a data warehouse for ecologists is not restricted to a specific application.

The general benefit of a data warehouse to an ecologist is that it provides a tool which can be used to easily explore the combined results of multiple related research projects. Attributes of interest can be extracted and further explored with data mining software or statistical packages in order to discover patterns that suggest potential ecological relationships which in turn can be the focus of future research. The query in Appendix B shows the relative simplicity of querying the data warehouse for attributes from the field data (latitude), derived from GIS analysis (water accumulation), aggregated from the field data (mean water depth) and calculated (fuzzy member ships for Shallow, Medium and Deep water depths). While there are many reasons for ecologists to move their data to a data warehouse, there are considerations to be made before beginning that project.

Creating a data warehouse is a significant undertaking, requiring an appreciable investment of time and resources. Much of this time is taken with cleaning the data, so is productive even if a data warehouse is not desired. Deciding which attributes are stored in a data warehouse is also a serious consideration, especially when aggregate values are to be stored instead of raw values. In this research fuzzy memberships for water depth was stored in the data warehouse, but the predictive power of these attributes was not tested so the arithmetic mean was also included. Likewise, as the field of bioacoustics evolves, PAA/SAX may not prove to be the best technique to reduce bioacoustic recordings. Here again, a different reduction can be derived and added to the data warehouse. These decisions must be informed by a thorough understanding of the data and the ways in which the data are to be used.

The next steps in this research are to continue development of the DSS, adding the reference library of acoustic and context attributes of candidate species, developing data

mining algorithms to match patterns between the acoustic recording and context to the reference library and creating a user interface. Assessment should be made on the effectiveness of the DSS to reduce processing time of bioacoustic recordings and any effect on accuracy. A comparison should also be made between the predictive power of fuzzy water depth memberships and mean water depth as measures of habitat preference. As well, further evaluation should be made of the effectiveness of PAA/SAX reduction of the bioacoustic recordings to be used for isolating and differentiating acoustic events.

This thesis demonstrates that moving ecological data into a data warehouse is possible and suggests many ways in which this can be an advantage to ecological research. The specific application of the data warehouse, to develops a DSS for bioacoustic processing, is something this researcher hopes to continue to develop.

Appendix A

Oversized tables and figures referenced in the thesis.

Table A1: Habitat values recorded by field staff which do not match standard DUEWC category abbreviations. The total of each category is shown at the bottom of the table as well as the number of entries which could not be matched.

Value Entered	Occurrence	Transposed Error	Partial Class Error	Multiple Class Error	Novel Entry Error	Omitted from DSS
?	1				✓	✓
BETULA	1				✓	✓
BHS	1					
BOG	2					
BOG UPLAND	2			✓		✓
BPR?	1				✓	✓
BURN	1				✓	
CATT	2				✓	✓
COMPRESSOR	3				✓	
CUTBLOCK	3				✓	
cutline	1				✓	
CUTLINE	51				✓	
FR6	1		✓		✓	

continued on next page

Table A1: Habitat values recorded by field staff which do not match standard DUEWC category abbreviations. The total of each category is shown at the bottom of the table as well as the number of entries which could not be matched.

Value Entered	Occurrence	Transposed Error	Partial Class Error	Multiple Class Error	Novel Entry Error	Omitted from DSS
FRT + FRS	2			✓		✓
FSR	2					
GFEN	2		✓			
GRAM FEN	14		✓			
GRAMINOID	2		✓			
GRAMINOID FEN	2		✓			
HIGHWAY	3				✓	
HWY	5				✓	
ILLEGIBLE	2				✓	✓
LAKE	1				✓	
NV	8				✓	✓
OF	4					✓
OPEN FEN	4			✓		✓
OPWA	11					
PATTERNED FEN	3				✓	✓
QUAD	2				✓	
RFG	2	✓				
ROAD	3				✓	
SEISMIC	20				✓	
SEISMIC LINE	1				✓	
SHRUBBY + TREED	2			✓		✓
SHRUBBY FEN	2					

continued on next page

Table A1: Habitat values recorded by field staff which do not match standard DUEWC category abbreviations. The total of each category is shown at the bottom of the table as well as the number of entries which could not be matched.

Value Entered	Occurrence	Transposed Error	Partial Class Error	Multiple Class Error	Novel Entry Error	Omitted from DSS
SHWBBY + GRAM FEN	2			✓		✓
SHWBBY + GRAMINOID FEN	2			✓		✓
SLOUGH	1				✓	✓
SNOWMOBILE TRAIL	1				✓	
SRH	2	✓				
STRING	2				✓	✓
TAMARACK	2		✓			
TF	1		✓			
TREED FEN	24		✓			
UNC	2	✓				
UNKNOWN	2				✓	✓
WELLPAD	2				✓	
null	19					✓
Totals:		6	48	14	120	62

Table A2: Translation of non-standard DUEWC habitat classifications to the three hierarchical levels of the classifications system, employing additional classes when necessary (shown in italics).

Value Entered	Auxillary Codes	Broad Cover Group	Major Wetland Group	Minor Wetland Group
BHS	BSH	Wetland	Bog	Shrubby
BOG	BOG	Wetland	Bog	*Undefined*
BURN	_DN	*Undefined*	*Undefined*	*Natural Disturbance*
COMPRESSOR	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
CUTBLOCK	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
CUTLINE	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
FR6	FR_	Wetland	Rich Fen	*Undefined*
FSR	FRS	Wetland	Rich Fen	Shrubby
GFEN	F_G	Wetland	*Undefined Fen*	Graminoid
GRAM FEN	F_G	Wetland	*Undefined Fen*	Graminoid
GRAMINOID	__G	*Undefined*	*Undefined*	Graminoid
GRAMINOID FEN	F_G	Wetland	*Undefined Fen*	Graminoid
HIGHWAY	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
HWY	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
LAKE	WAT	Wetland	Water	Open
MEADOW	_MD	*Undefined*	*Undefined*	Meadow
OPWA	WAT	Wetland	Water	Open
QUAD	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*

continued on next page

Table A2: Translation of non-standard DUEWC habitat classifications to the three hierarchical levels of the classifications system, employing additional classes when necessary (shown in italics).

Value Entered	Auxillary Codes	Broad Cover Group	Major Wetland Group	Minor Wetland Group
RFG	FRG	Wetland	Rich Fen	Graminoid
ROAD	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
SEISMIC	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
SEISMIC LINE	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
SHRUBBY FEN	F_S	Wetland	*Undefined Fen*	Shrubby
SNOWMOBILE TRAIL	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
SRH	SHR	Wetland	Swamp	Hardwood
TAMARACK	_TM	*Undefined*	*Undefined*	Tamarack
TF	F_T	Wetland	*Undefined Fen*	Tamarack
TREED FEN	F_T	Wetland	*Undefined Fen*	Tamarack
UNC	UCN	Upland	Upland	Treed
WELLPAD	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
cutline	_DA	*Undefined*	*Undefined*	*Anthropogenic Disturbance*
meadow	_MD	*Undefined*	*Undefined*	Meadow

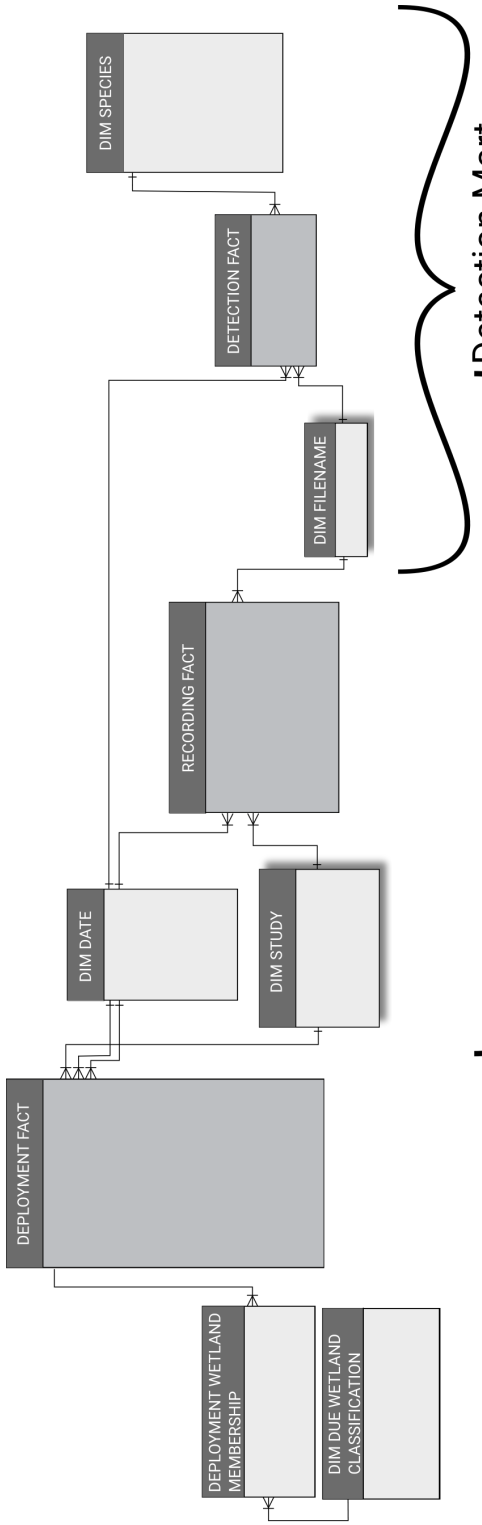


Figure A1: All data marts combined to form the complete data warehouse.

Appendix B

Sample query and results from the data warehouse.

Box B1: Query written for the data warehouse created in this thesis for the purpose of extracting some of the context data for recordings where Yellow Rails (species id = 2333) had been identified.

```
SELECT
    DETF.ID_DIM_SPECIES,
    DEPF.LATITUDE,
    DEPF.WATER_ACCUM,
    DEPF.CANVEC_WATERBODY_PROP,
    DEPF.CANVEC_WETLAND_PROP,
    DEPF.TOTAL_HORIZONTAL_COVER,
    DEPF.MEAN_DEPTH,
    DEPF.MF_MEAN_SHALLOW,
    DEPF.MF_MEAN_MEDIUM,
    DEPF.MF_MEAN_DEEP
FROM DEPLOYMENT_FACT DEPF
INNER JOIN RECORDING_FACT RECF
ON DEPF.ID_DIM_STUDY = RECF.ID_DIM_STUDY
INNER JOIN DETECTION_FACT DETF
ON RECF.ID_DIM_FILENAME = DETF.ID_DIM_FILENAME
WHERE DETF.ID_DIM_SPECIES = 2333;
```

Table B1: The first ten records returned from the query (above) executed on the data warehouse created in this thesis.

ID_DIM_SPECIES	LATITUDE	WATER_ACCUM	CANVEC_WATERBODY_PROP	CANVEC_WETLAND_PROP	TOTAL_HORIZONTAL_COVER	MEAN_DEPTH	MF_MEAN_SHALLOW	MF_MEAN_MEDIUM	MF_MEAN_DEEP
2333	55.05376	114.1989	(null)	1	(null)	(null)	(null)	(null)	(null)
2333	56.11446	45.19996	(null)	(null)	(null)	(null)	(null)	(null)	(null)
2333	55.79884	11.90459	(null)	(null)	(null)	(null)	(null)	(null)	(null)
2333	54.62604	109.2803	0.82	0.18	(null)	(null)	(null)	(null)	(null)
2333	54.62604	109.2803	0.82	0.18	(null)	(null)	(null)	(null)	(null)
2333	54.55645	75.72627	0.03	0.74	0.391	0.95	0.92	0.08	0
2333	54.56444	174.6789	(null)	0.91	0.525	5.62	0.53	0.43	0.05
2333	57.48545	6.753703	0.3	0.02	(null)	7.29	0.86	0	0.14
2333	57.45776	315.0177	(null)	1	(null)	(null)	(null)	(null)	(null)
2333	57.43589	1.600749	(null)	1	(null)	5.71	0.4	0.58	0.02

Literature Cited

Agranat, I. (2009). Automatically identifying animal species from their vocalizations. In *Fifth International Conference on Bio-Acoustics*, pages 1–22, Concord, Massachusetts, USA. Wildlife Acoustics, Inc.

Azevedo, A. and Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining*, (January):182–185.

Bagnall, A., Lines, J., Bostrom, A. G., Large, J., and Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660.

Bagnall, A., Lines, J., Hills, J., and Bostrom, A. G. (2015). Time-series classification with COTE: The collective of transformation-based ensembles. *2016 IEEE 32nd International Conference on Data Engineering, ICDE 2016*, 27(9):1548–1549.

Beeby, A. and Brennan, A. (2007). *First ecology: ecological principles and environmental issues*. Oxford University Press, 3rd edition.

Bivand, R. and Lewin-Koh, N. (2016). Maptools: Tools for reading and handling spatial objects [computer software]. Available from http://cran.r-project.org.

Bostrom, A. G. and Bagnall, A. (2015). Binary shapelet transform for multiclass time series classification. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 257–269. Springer.

Brandes, T. S. (2008). Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International*, 18(S1).

Breslin, M. (2004). Data warehousing battle of the giants : Comparing the basics of the Kimball and Inmon models. *Business Intelligence Journal*, pages 6–20.

British Columbia Ministry of Forests (1995). *Summary of British Columbia forest inventory statistics by land administration class.*, volume 1. Victoria, B.C.

Chambert, T., Waddle, J. H., Miller, D. A., Walls, S. C., and Nichols, J. D. (2018). A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Methods in Ecology and Evolution*, 9(3):560–570.

Charif, R. A. and Pitzrick, M. (2008). Automated detection of Cerulean Warbler songs using XBAT data template detector software: preliminary report. Technical report, Cornell Lab of Ornithology, Bioacoustics Research Program, Ithaca, NY.

Darras, K., Batáry, P., Furnas, B., Fitriawan, I., Mulyani, Y., and Tscharntke, T. (2017). Autonomous bird sound recording outperforms direct human observation: Synthesis and new evidence. *bioRxiv*, pages 1–37.

Ducks Unlimited (2015). *Field Guide Boreal Wetland Classes*. Number January. First edit edition.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. Oxford University Press.

El-Sappagh, S. H. A., Hendawi, A. M. A., and El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University - Computer and Information Sciences*, 23(2):91–104.

Fagerlund, S. (2004). *Automatic recognition of bird species by their sounds*. Master of science in technology, Helsinki University of Technology.

Fayyad, U. (1997). Data mining and knowledge discovery in databases: implications for scientific databases. In *Proceedings. Ninth International Conference on Scientific and Statistical Database Management (Cat. No.97TB100150)*, pages 2–11.

Friederichs, K. (1958). A definition of ecology and some thoughts about basic concepts. *Ecology*, 39(1):154–159.

GRASS Development Team (2015). Geographic Resources Analysis Support System (GRASS GIS) [computer software]. Open Source Geospatial Foundation.

Gregory, R. D. and Strien, A. V. (2010). Wild bird indicators : using composite population trends of birds as measures of environmental health. *Ornithological Science*, 22:3–22.

Gupta, V. R. (1997). An introduction to data warehousing. *System Services Corporation, Chicago, Illinois*, 11(August).

Gutiérrez-Estrada, J. C., Pulido-Calvo, I., and Bilton, D. T. (2013). Consistency of fuzzy rules in an ecological context. *Ecological Modelling*, 251:187–198.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S., and Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3):156–162.

Hills, J., Lines, J., Baranauskas, E., Mapp, J., and Bagnall, A. (2014). Classification of time series by shapelet transformation. *Data Mining and Knowledge Discovery*, 28(4):851–881.

Hillyer, M. (2005). An introduction to database normalization. Technical report, MySQL AB.

Hinton, M. (2006). *Introducing Information Management.* Routledge.

Hutto, R. and J. Stutzman, R. (2009). Humans versus autonomous recording units: A comparison of point-count results. *Journal of Field Ornithology*, 80:387–398.

Inmon, W. H. (1996). The data warehouse and data mining. *Communications of the ACM*, 39(11):49–50.

Inmon, W. H. and Kelley, C. (1993). *Rdb-VMS: Developing a data warehouse.* John Wiley & Sons, Inc.

Jones, M. B., Schildhauer, M. P., Reichman, O. J., and Bowers, S. (2006). The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics*, 37(1):519–544.

Kasten, E. P., Gage, S. H., Fox, J., and Joo, W. (2012). The remote environmental assessment laboratory's acoustic library: An archive for studying soundscape ecology. *Ecological Informatics*, 12:50–67.

Kasten, E. P. and McKinley, P. K. (2007). MESO: Supporting online decision making in autonomic computing systems. 19(4):485–499.

Kasten, E. P., McKinley, P. K., and Gage, S. H. (2007). Automated ensemble extraction and analysis of acoustic data streams. In *Proceedings - International Conference on Distributed Computing Systems*, volume 1, Toronto.

Kate, R. J. (2016). Using dynamic time warping distances as features for improved time series classification. *Data Mining and Knowledge Discovery*, 30(2):283–312.

Keen, S., Ross, J. C., Griffiths, E. T., Lanzone, M., and Farnsworth, A. (2014). A comparison of similarity-based approaches in the classification of flight calls of four species of North American wood-warblers (Parulidae). *Ecological Informatics*, 21:25–33.

Keogh, E., Chakrabarti, K., Pazzani, M., and Mehrotra, S. (2000). Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and Information Systems*, 3(3):263–286.

Kimball, R. (1996). The data warehouse toolkit: practical techniques for building dimensional data warehouse. *NY: John Willey & Sons*, 248(4).

Kimball, R. and Caserta, J. (2004). *The data warehouse ETL toolkit*, volume 1. Wiley Publishing, Inc., Indianapolis, IN.

Kimball, R., Ross, M., Thorthwaite, W., Becker, B., and J, M. (2008). *The data warehouse lifecycle toolkit, 2nd edition.* John Wiley & Sons, Inc.

Kimball, R., Ross, M., Wiley, J., and Anisimov, A. A. (2002). *The data warehouse toolkit: The complete guide to dimensional modeling (2nd edition)*, volume 464. John Wiley & Sons, 2 edition.

Kirschel, A. N. G., Earl, D. A., Yao, Y., Escobar, I., Vilches, E., Vallejo, E. E., and Taylor, C. E. (2009). Using songs to identify individual mexian antthrush Formicarius monilinger: comparison of four classification methods. *Bioacoustics*, 19:1–20.

La, V. T. and Nudds, T. D. (2016). Estimation of avian species richness: Biases in morning surveys and efficient sampling from acoustic recordings. *Ecosphere*, 7(4):1–13.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 2–11.

Lin, J., Khade, R., and Li, Y. (2012). Rotation-invariant similarity in time series using bag-of-patterns representation. *Journal of Intelligent Information Systems*, 39(2):287–315.

Lin, J., Lonardi, J., and Patel, P. (2002). Finding motifs in time series. In *Proc. of the 2nd Workshop on Temporal Data Mining*, pages 53–68.

Madin, J., Bowers, S., and Schildhauer, M. (2007). An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3):279–296.

Mannan, R. W. (1984). Habitat use by Hammond's flycatchers in old-growth forests, northeastern Oregon. *Murrelet*, 65(3):84–86.

Marchini, A., Facchinetti, T., and Mistri, M. (2009). F-IND: A framework to design fuzzy indices of environmental conditions. *Ecological Indicators*, 9(3):485–496.

Ministry of Natrural Resources (1997). Canadian Digital Elevation Model (CDEM).

Moody, D. and Kortink, M. A. (2000). From enterprise models to dimensional models: A methodology for data darehouse and data mart design. *Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000)*, 2000:5–16.

Natural Resources Canada (2014). CanVec+ Feature Catalogue. Technical report, Natural Resources Canada, Sherbrooke, QC.

North American Classification Committee (2014). AOU Checklist of North and Middle American Birds (7th Edition and Supplements). Technical report, American Ornithologists' Union.

Obrist, M., Pavan, G., Sueur, J., and Riede, K. (2010). Bioacoustics approaches in biodiversity inventories. *Abc Taxa*, pages 68–99.

Palmer, M. A., Bernhardt, E. S., Chornesky, E. A., Collins, S. L., Dobson, A. P., Duke, C. S., Gold, B. D., Jacobson, R. B., Kingsland, S. E., Kranz, R. H., Mappin, M. J., Martinez, M. L., Micheli, F., Morse, J. L., Pace, M. L., Pascual, M., Palumbi, S. S., Reichman, O. J., Townsend, A. R., Turner, M. G., Frontiers, S., Feb, S. F., Chornesky, E. A., Collins, S. L., Dobson, A. P., Palmer, M. A., Bernhardt, E. S., Duke, S., Gold, B. D., Jacobson, R. B., Kingsland, S. E., Kranz, R. H., Mappin, M. J., Martinez, M. L., Micheli, F., Morse, J. L., Pace, M. L., Pascual, M., Stephen, S., Townsend, A. R., and Turner, M. G. (2005). Ecological Science and Sustainability for the 21st Century. *Frontiers in Ecology and the Environment*, 3(1):4–11.

Pouw, F. and Kwiatkowska, M. (2013). An overview of fuzzy-logic based approaches to ecology: Addressing uncertainty. In *2013 Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, pages 540–545.

R Core Team (2017). R: A language and environment for statistical computing.

Rakthanmanon, T. and Keogh, E. (2013). Fast shapelets: A scalable algorithm for discovering time series shapelets. *Proceedings of the 2013 SIAM International Conference on Data Mining*, 31(5):668–676.

Regan, H. M., Colyvan, M., and Burgman, M. A. (2012). A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications*, 12(2):618–628.

Rempel, R. S., Hobson, K. A., Holborn, G., Van Wilgenburg, S. L., and Elliott, J. (2005). Bioacoustic monitoring of forest songbirds: interpreter variability and effects of configuration and digital processing methods in the laboratory. *Journal of Field Ornithology*, 76(1):1–11.

Sauer, J. R., Link, W. A., Fallon, J. E., Pardieck, K. L., and Ziolkowski, D. J. (2013). The North American Breeding Bird Survey 1966-2011: Summary analysis and species accounts. *North American Fauna*, 79(79):1–32.

Sedgwick, J. A. (1975). *A comparative study of the breeding biology of Hammond's (Empidonax hammondii) and Dusky (Empidonax oberholseri) flycatchers*. PhD thesis, University of Montana.

Sen, A. and Sinha, A. P. (2005). A comparison of data warehousing methodologies. *Communications of the ACM*, 48(3):79–84.

Senin, P. (2016). jmotif: Time series analysis toolkit based on symbolic aggregate dicretization, i.e. SAX.

Senin, P. and Malinchik, S. (2013). SAX-VSM: Interpretable time series classification using sax and vector space model. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 298(0704):1175–1180.

Shonfield, J. and Bayne, E. M. (2017b). Autonomous recording units in avian ecological research: current use and future applications. *Avian Conservation and Ecology*, 12(1):art14.

Suh, S. C. (2012). *Practical applications of data mining*. Jones & Bartlet Learning, Mississauga.

Thuraisingham, B. (1997). *Data management systems: Evolution and interoperation*. CRC Press.

Towsey, M. and Planitz, B. (2011). Technical Report: Acoustic analysis of the natural environment. Technical report.

Towsey, M., Planitz, B., Nantes, A., Wimmer, J., and Roe, P. (2012). A toolbox for animal call recognition. *Bioacoustics*, 21(2):107–125.

Truskinger, A., Towsey, M., and Roe, P. (2015). Decision support for the efficient annotation of bioacoustic events. *Ecological Informatics*, 25(January):14–21.

Truskinger, A., Yang, H., Wimmer, J., Zhang, J., Williamson, I., and Roe, P. (2011). Large scale participatory acoustic sensor data analysis: Tools and reputation models to dnhance dffectiveness. In *Proceedings - 2011 7th IEEE International Conference on eScience, eScience 2011*, pages 150–157.

Wimmer, J., Towsey, M., Planitz, B., Williamson, I., and Roe, P. (2013). Analysing environmental acoustic data through collaboration and automation. *Future Generation Computer Systems*, 29(2):560–568.

Yi, B.-K. and Faloutsos, C. (2000). Fast time sequence indexing for arbitrary Lp norms. In *VLDB*, volume 385, page 99. Citeseer.

Zadeh, L. (1965). Fuzzy Sets. *Information and Control*, 8(3):338–&.

Zhang, L., Towsey, M., Zhang, J., and Roe, P. (2016). Classifying and ranking audio clips to support bird species richness surveys. *Ecological Informatics*, 34:108–116.